Cross-Entropy Induces Linear Representations

Su Hyeong Lee [1]

Abstract

Why do some internal concepts in a neural network become readily accessible to simple linear probes, while others remain difficult to extract? In this paper, we show that when predictive structure is expressed through finitely many feature coordinates and compressed through a low-dimensional bottleneck, next-token cross-entropy has an implicit bias toward pre-linearization. The model preferentially preserves interaction directions that are both statistically common in the data and highly useful for prediction. In the simplest orthogonal setting, this yields an explicit commonness-times-usefulness law predicting which interactions become linearly readable first. In the general case, however, the relevant object is not an individual named feature, but a basis-invariant whitened subspace of interaction space. We test these predictions in controlled synthetic settings, trained transformers, and pretrained language models. The directions singled out by the theory are the ones that become most linearly readable across all three settings. In our controlled experiments, they are also the most effective and selective directions for steering model behavior. The complete codebase for reproducing all experiments is available in the **project repository**.

1. Introduction

Some internal interactions in a trained network become linearly readable from intermediate activations long before others. Linear probes make this visible [1–3], and recent work on language models has connected linear representability to pretraining frequency [4], latent-variable models of next-token prediction [5], and sparse-plus-low-rank implicit-geometry accounts [6]. The training objective, however, is next-token cross-entropy, not an explicit linearity penalty. This leaves a more specific question open: which interaction directions does cross-entropy itself prefer when predictive structure has to be compressed through a low-dimensional bottleneck?

Frequency-based accounts suggest a natural explanation for why some concepts become linearly readable: concepts that appear often in pretraining data are encountered more often, and may therefore be easier for the model to represent [4]. But frequency alone is not a satisfactory selection rule. It does not identify the mathematical object optimized by next-token cross-entropy when predictive structure must pass through a low-dimensional bottleneck, nor does it explain why the retained predictive structure should appear as a linear subspace accessible to probes. When the model cannot preserve every interaction, precisely which directions does it retain, and in which metric? Why should the retained structure be linear? Our answer identifies the selection rule at the level of the training objective and expresses it as a concrete operator. We show that cross-entropy selects a whitened interaction-to-logit subspace, yielding a basis-invariant rule that predicts linear readability, feature selection, and local steering. In this sense, linearization is not an incidental property of trained models, but an implicit bias of cross-entropy under representation bottlenecks. Classical low-rank frameworks approach related questions from another direction. Appendix A discusses the relationship in detail.

Fix a dictionary of candidate interaction features, $\phi : X \to \mathbb{R}^m$, where each coordinate $\phi_r(x)$ measures one interaction feature present in the context x. This paper shows that, under this fixed interaction dictionary and a width-k representation bottleneck, bounded-logit next-token cross-entropy is uniformly comparable – via an explicit softmax-curvature sandwich – to low-rank approximation of the whitened interaction-to-logit operator

$$T = C\Sigma^{1/2}.$$

The induced Frobenius surrogate is solved exactly by the top-k right singular subspace of T, and under a spectral gap and a small bounded-logit sandwich distortion, the true cross-entropy optimum recovers the same subspace up to an explicit principal-angle bound. Here, $C\phi$ is the L_2 projection of the teacher logits onto the dictionary-generated family, and $\Sigma = \mathbb{E}[\phi(X)\phi(X)^\top]$ is the uncentered feature second moment. Under arbitrary dictionary reparameteriza-

[1]Department of Statistics, University of Chicago. Correspondence to: Su Hyeong Lee <sulee@uchicago.edu>.

Accepted at the ICML 2026 Workshop on Connecting Low-rank Representations in AI, Seoul, South Korea. Copyright 2026 by the author(s).

tions, the selected object is a basis-invariant whitened subspace rather than a feature list. At middle layers, the same geometry reappears as the pullback operator $T_\ell = J_\ell B \Sigma^{1/2}$, whose top singular directions predict both later linear readability and selective steering.

In the special case where the dictionary coordinates and their logit effects are orthogonal, the subspace rule reduces to a feature-wise product law, $d_r = \nu_r \|c_r\|_2^2$. The score ranks each feature by two quantities: the uncentered second moment $\nu_r = \mathbb{E}[\phi_r(X)^2]$ – which for centered indicators reduces to prevalence $p_r(1-p_r)$ – and the squared magnitude $\|c_r\|_2^2$ of its logit-space coefficient. Thus, in orthogonal coordinates, cross-entropy preferentially preserves interaction directions that are both statistically common in the data and logit-wise useful for prediction, and makes those directions linearly available through the bottleneck. The product form goes beyond frequency-based accounts suggested by prior work: it predicts crossovers that neither commonness nor logit-wise usefulness alone can capture. A rare but highly predictive feature outranks a common but barely predictive one whenever the product d_r is larger.

Outside the orthogonal regime, individual named features lose their status as canonical objects. The invariant selected by cross-entropy is a k-dimensional subspace of whitened feature coordinates, preserved under any invertible change of dictionary basis, whereas named top-k feature lists are basis dependent. At a middle layer ℓ, the same geometry reappears as the pulled-back operator $T_\ell = J_\ell B \Sigma^{1/2}$, where J_ℓ is the downstream Jacobian from the layer to the logits and B maps features to the hidden state. Its top singular directions are first-order optimal for local steering per unit whitened-feature perturbation. Under a hidden-Euclidean-norm budget, the optimizer is a generalized-eigenvalue analogue (Section 4, Appendix K.5).

We use "pre-linearization" for the empirical phenomenon that some interaction features become linearly readable from intermediate activations during training while others remain entangled. The theory above identifies which k-dimensional subspace cross-entropy selects. The experiments test whether empirical probes recover that subspace.

We test seven empirical predictions across settings of increasing realism. In synthetic teacher–student models, we test the product-rule crossover law, basis invariance under dictionary reparameterization, and a matched-pair intervention showing that the product score, rather than commonness or usefulness alone, controls within-pair emergence order. In a controlled transformer, we test whether the predicted subspace is preserved from early to late training, agrees with the empirically readable subspace, and yields the pulled-back local-control law. Finally, in pretrained Pythia models, we test whether the same operator prospectively forecasts the task-relevant subspace on subject–verb agreement.

2. Mathematical setup

Let (X, μ) be a probability space of contexts and let $G = \{z \in \mathbb{R}^n : \mathbf{1}^\top z = 0\}$ be the zero-sum logit gauge. The teacher next-token distribution is represented by logits $\eta_* : X \to G$ with $p_*(\cdot \mid x) = \mathrm{softmax}(\eta_*(x))$ and $\eta_* \in L_2(\mu; G)$.

Fix an interaction dictionary $\phi = (\phi_1, \ldots, \phi_m) : X \to \mathbb{R}^m$ with second-moment matrix

$$\Sigma := \mathbb{E}[\phi(X)\phi(X)^\top] \succ 0.$$

Let $\mathcal{F} = \{M\phi(\cdot) : M \in \mathbb{R}^{n\times m}, \mathbf{1}^\top M = 0\}$ be the corresponding dictionary-generated logit family. Writing the $L_2(\mu; G)$ projection of η_* onto $\mathcal{F}$ as

$$P_{\mathcal{F}}\eta_* = C\phi, \qquad r := \eta_* - C\phi,$$

we define the whitened interaction-to-logit operator

$$T := C\Sigma^{1/2} \in \mathbb{R}^{n\times m}.$$

A width-k dictionary bottleneck is any student of the form

$$\widehat{\eta}_M(x) = M\phi(x), \qquad \mathrm{rank}(M) \le k, \qquad \mathbf{1}^\top M = 0,$$

equivalently any factorization $\widehat{\eta}(x) = Uh(x)$ with $h(x) \in \mathbb{R}^k$ linear in $\phi(x)$. Its population risk and the Bayes risk are

$$\mathcal{R}(M) := \mathbb{E}_{X,Y\sim p_*}[-\log \mathrm{softmax}(M\phi(X))_Y],$$
$$\mathcal{R}_* := \mathbb{E}_{X,Y\sim p_*}[-\log p_*(Y \mid X)].$$

Finally, for $B > 0$ let

$$K_B := \{z \in G : \|z\|_\infty \le B\}.$$

On K_B, the softmax Hessian is uniformly bounded above and below on G by constants $0 < m_B \le M_B < \infty$; Appendix D.3 makes these constants explicit.

3. Final-layer spectral theory

Theorem 3.1 (Bounded-logit reduction)**.** *Assume $\eta_*(x) \in K_B$ and $M\phi(x) \in K_B$ for μ-almost every x. Then every admissible width-k student satisfies*

$$\frac{m_B}{2}\Big(\|r\|^2_{L_2(\mu;G)} + \mathcal{E}(M)\Big) \le \mathcal{R}(M) - \mathcal{R}_* \le \frac{M_B}{2}\Big(\|r\|^2_{L_2(\mu;G)} + \mathcal{E}(M)\Big),$$

where $r = \eta_* - C\phi$ is the projection residual. The residual term $\|r\|^2_{L_2(\mu;G)}$ is independent of M and captures the dictionary-mismatch error from restricting the teacher logits to $\mathcal{F}$. Thus, for a fixed dictionary, the rank-k student can only control

$$\mathcal{E}(M) := \left\|(C - M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

Equivalently, bounded-logit cross-entropy is uniformly comparable to low-rank approximation of the whitened interaction-to-logit operator $T = C\Sigma^{1/2}$.

Theorem 3.1 provides the reduction from next-token cross-entropy to pre-linearization. Once the dictionary has been fixed, the cross-entropy optimum is governed by the same object that governs best rank-k approximation of T. The residual term $\|r\|^2_{L_2(\mu;G)}$ measures dictionary mismatch, and the matrix term measures what the bottleneck can or cannot keep. Because the surrogate is Frobenius, the Eckart–Young–Mirsky theorem identifies the width-k optimum explicitly.

Theorem 3.2 (Exact width-k solution)**.** *Let*

$$T = U \operatorname{Diag}(s_1, \dots, s_q, 0, \dots, 0) V^\top,$$

with $s_1 \geq s_2 \geq \cdots \geq s_q > 0$, be an SVD of $T = C\Sigma^{1/2}$. Then

$$\min_{\operatorname{rank}(M) \leq k} \mathcal{E}(M) = \sum_{i=k+1}^{q} s_i^2,$$

and an optimizer is

$$M_k = U_k \operatorname{Diag}(s_1, \dots, s_k) V_k^\top \Sigma^{-1/2}.$$

Since $\mathbf{1}^\top T = 0$, every nonzero left singular vector of T lies in the zero-sum gauge G, and this optimizer satisfies $\mathbf{1}^\top M_k = 0$. Hence the optimal width-k bottleneck keeps a top-k right singular subspace of T; when $s_k > s_{k+1}$ this subspace is unique.

The optimizer in Theorem 3.2 is stated in terms of the right singular vectors V_k of T, which depend on the chosen dictionary. Under a reparameterization $\phi \to A\phi$, the coordinates V_k change; however, the subspace they span, viewed in whitened feature coordinates, does not.

Theorem 3.3 (Basis invariance of the selected whitened subspace)**.** *Let $\phi' = A\phi$ for some invertible matrix $A \in \mathbb{R}^{m\times m}$, and define*

$$\Sigma' := \mathbb{E}[\phi'(X)\phi'(X)^\top] = A\Sigma A^\top.$$

Let C' satisfy $C'\phi' = C\phi$ and set $T' := C'{\Sigma'}^{1/2}$. Then

$$Q := {\Sigma'}^{-1/2} A \Sigma^{1/2}$$

is orthogonal and

$$T' = TQ^\top.$$

Consequently T and T' have the same singular values. If V_k spans the top-k right singular subspace of T, then QV_k spans the top-k right singular subspace of T', and the readable scalar space

$$\operatorname{span}\{v^\top \Sigma^{-1/2}\phi(\cdot) : v \in \operatorname{range}(V_k)\}$$

is exactly equal to

$$\operatorname{span}\{{v'}^\top {\Sigma'}^{-1/2}\phi'(\cdot) : v' \in \operatorname{range}(QV_k)\}.$$

Thus the canonical object is a basis-invariant whitened interaction subspace, not a basis-dependent coordinate list.

A proof is given in Appendix H. Theorem 3.3 identifies the whitened interaction subspace, rather than any list of named coordinates, as the invariant of the bottleneck problem. A named-feature ranking recovers this subspace only in the special case where it aligns with the dictionary coordinate axes.

Corollary 3.4 (Orthogonal-coordinate ranking)**.** *Suppose the dictionary coordinates are orthogonal in $L_2(\mu)$ with variances $\nu_r = \mathbb{E}[\phi_r(X)^2]$, and the corresponding output directions c_r (the columns of C) are pairwise orthogonal in G. Then the optimal width-k student keeps the k coordinates with largest score*

$$d_r := \nu_r \|c_r\|_2^2.$$

For centered indicator interactions, $\nu_r = p_r(1-p_r)$, so in the rare-event regime the score is proportional to prevalence times predictive usefulness.

A proof is given in Appendix G. Corollary 3.4 formalizes, in the orthogonal-coordinate case, the principle that common and useful features are pre-linearized first. The commonness factor is ν_r, and the usefulness factor is the logit-wise predictive strength $\|c_r\|_2^2$. The product form predicts crossovers that neither commonness-only nor usefulness-only heuristics can capture.

Two related results are deferred to the appendix for space. Appendix F–H shows that, under a spectral gap $s_k(T) > s_{k+1}(T)$ and a small bounded-logit sandwich gap $M_B/m_B - 1$, the true cross-entropy optimum recovers the same top-k whitened subspace as the surrogate, with a principal-angle bound explicit in both gaps. Appendix I analyzes aligned balanced gradient flow on the factorized surrogate and shows that modes with larger singular values are amplified earlier. In this stylized dynamics model, the emergence order follows the singular-value order $s_1 \geq s_2 \geq \cdots$. Thus, the spectral object predicts not only which directions are selected at optimum, but also which directions should become linearly readable first in the surrogate dynamics.

4. Pullback to middle layers and local control

Fix a layer ℓ with hidden state $h_\ell(x) \in \mathbb{R}^d$. We model the teacher state as $h_*(x) = B\phi(x)$ and a width-k student as $h_M(x) = M\phi(x)$. Let $F_\ell : U \to G$ be the downstream map from layer ℓ to the logits. If F_ℓ is affine, $F_\ell(u) = J_\ell u + b_\ell$, the $L_2(\mu; G)$ excess risk at layer ℓ is sandwiched between constants times $\|J_\ell(B-M)\Sigma^{1/2}\|_F^2$ (Appendix K),

so the layer-ℓ pulled-back operator

$$T_\ell := J_\ell B \Sigma^{1/2}$$

governs the layer-ℓ compression problem; the optimal width-k representation keeps the top singular directions of T_ℓ. Under only locally linear tails (e.g., transformer tails on the relevant task region), the same reduction holds up to a distortion factor $(\frac{1+\delta_\ell/\alpha_\ell}{1-\delta_\ell/\alpha_\ell})^2$, where δ_ℓ and α_ℓ quantify local non-linearity and injectivity on the task subspace. The statement requires $\delta_\ell < \alpha_\ell$. The residual-stream specialization gives an explicit telescoping expression $J_\ell = W\,\Pi_\ell$ for affine residual updates.

The local-control counterpart parameterizes task-subspace perturbations by whitened feature coordinates. Among k-dimensional subspaces $S \subset \mathbb{R}^m$ of the whitened-feature space and unit-norm directions $u \in S$, the top singular directions of T_ℓ maximize the first-order logit energy per unit-norm whitened-feature perturbation:

$$\max_{\dim S = k} \mathbb{E}_{u \sim \mathrm{Unif}(S \cap \mathbb{S}^{m-1})} \|T_\ell u\|_2^2 \;=\; \frac{1}{k}\sum_{i=1}^{k} s_i(T_\ell)^2,$$

attained on the top-k right singular subspace. The associated hidden perturbation is $\delta h = B\Sigma^{1/2}u$. Thus the singular-vector statement is exactly optimal per unit whitened-feature-coordinate norm; if the intervention budget is the hidden Euclidean norm $\|\delta h\|_2$, the corresponding optimizer is a generalized eigenproblem (Appendix K.5). Unbounded perturbations remain outside the local-control regime. Appendix K gives the two reduction theorems, the residual corollary, and both local-control norm statements with proofs.

5. Experiments

Table 1 summarizes the seven quantitative predictions tested in this section, together with the corresponding experiment, primary metric, and falsifying outcome. We test these in three settings of increasing realism: orthogonal and correlated synthetic teacher–student models, a controlled transformer trained on a synthetic task with a known interaction dictionary, and pretrained Pythia language models on subject–verb agreement minimal pairs. Complete protocols, feature definitions, hyperparameters, and the causal-intervention predictions are in Appendix B and the supplementary material. All reported numbers use 8 seeds unless stated otherwise, with 95% bootstrap confidence intervals across seeds. We use a single emergence rule. A feature has emerged at step t if its held-out probe R^2 exceeds 0.8 for three consecutive checkpoints.

5.1. Product-rule crossovers on orthogonal features

Using the fixed 24-feature set (24 features in three groups of 8: 4 tie pairs, 4 crossover pairs, and 4 control pairs, with the control pairs split as two same-usefulness and two same-commonness; complete specification in Appendix B), we trained direct-CE and surrogate rank-k students for $k \in \{1, 2, 4, 8, 12\}$ on 200,000 samples across 8 seeds. Because Gaussian features are unbounded, the bounded-logit sandwich of Theorem 3.1 is not a population guarantee for the original Gaussian law. For the finite empirical distribution used to plot Figure 2, we apply a single global teacher rescaling so that all evaluated teacher, trained rank-k student, and random rank-k student logits lie in K_B with $B = 2.5$; the sandwich then applies on that support.

Risk–surrogate coupling. Figure 2 pools the excess cross-entropy $\mathcal{R}(M) - \mathcal{R}_*$ against the surrogate error $\|T - M\Sigma^{1/2}\|_F^2$ across all $(k, \text{optimizer}, \text{seed})$ trained optima and 100 random rank-k matrices per configuration. The two quantities are approximately linearly related over the range plotted, consistent with the bounded-logit sandwich of Theorem 3.1.

Product-rule ranking accuracy on adversarial pairs. For each non-tie pair in any group, the three rankings (product $d = \nu\gamma^2$, commonness-only ν, usefulness-only γ^2) predict a winner, and we record whether each ranking's predicted winner has the higher probe R^2 at $k = 12$. Tie pairs, where the product scores of the two members are equal by construction, are excluded from the winner-prediction denominator and evaluated separately by the tie-gap metric below. When commonness-only or usefulness-only is itself indifferent on a pair because the pair gives both features the same value of that factor, the pair is counted as non-informative for that baseline. Under this rule, the product rule selects the higher-R^2 feature on $\mathbf{100}\%$ of 64 non-tie pairs. Commonness-only does so on 42.2% of the same pairs, and usefulness-only does so on 100% of the 48 informative pairs (Figure 1 left).

Tie pairs have smaller within-pair gaps. Tie pairs test the prediction of no strict ordering. For these pairs, the product scores of the two features are equal, so the product rule predicts no strict ordering. We therefore ask whether the two features become similarly readable after training, measured by the within-pair probe gap $|R_a^2 - R_b^2|$. At $k = 12$, the median gap is $\mathbf{0.248}$ for tie pairs, 0.509 for crossover pairs, and 0.863 for control pairs. Thus, tied-product pairs are roughly $2\times$ tighter than crossover pairs and $3.5\times$ tighter than control pairs (Figure 1, right). The within-tie gap is not zero, because the top-k rank-k selection at $k = 12$ is underdetermined when many tied features compete. Nevertheless, the monotone ordering from tie to crossover to control pairs is consistent with the pattern predicted by the product rule.

In a conjunction-feature replication (centered pairwise Bernoulli features on 12 primitives; protocol in Appendix B), the product rule achieves $\mathbf{82.6}\%$ accuracy

Table 1. Predictions tested in Section 5, the corresponding experiment, the primary reported metric with confidence interval, and the outcome that would not support the prediction.

Prediction	Experiment	Primary metric (reference)	Not support prediction if…
Product-rule ranking in orthogonal coordinates	Orthogonal Gaussian + conjunction benchmark (§5)	Adversarial-pair win rate 1.000 and 0.826 (non-tie pool)	Single-axis rule above product rule
Product rule is causal, not correlational	Matched-pair intervention (§5.3)	DiD $+1.94, +1.92$ at k=8 (8 seeds)	Single-axis rule saturates DiD in both families
Basis invariance of the selected subspace	Correlated mixing + transport (§5)	Transported top-k overlap ≈ 1.000	Named-coordinate list preserved under mixing
Early-to-late subspace preservation	Controlled transformer (§5.4)	$\Omega_4 = 0.74 \pm 0.07$	Overlap indistinguishable from the random-subspace baseline $k/m \approx 0.17$
Pullback subspace matches empirically readable subspace	Controlled transformer (§5.4)	$\Omega_4^{\text{pred-vs-read}} = 0.51 \pm 0.07$	Overlap indistinguishable from the random-subspace baseline
Top-k singular directions maximize the tested local-control metrics	Frozen-tail steering (§5.5)	Singular directions exceed probe, mean-difference, and random baselines on all three metrics at ℓ=2, k=4	Any baseline family exceeds singular
Pulled-back forecasts carry more task-specific signal than hidden-PCA at early checkpoints	Pythia SVA (§5.6)	Ratio $6.8\times$–$10.3\times$ at step 4000	Hidden-PCA matches or exceeds pulled-back at early checkpoints

on 63 non-tie pairs, versus 52.2% for commonness-only and 87.5% for usefulness-only (informative on 47 pairs). The product rule identifies the higher-R^2 feature in both crossover *and* control pair families in both dictionary variants, whereas commonness-only is inaccurate on the crossover family in both (0.0% Gaussian, 18.8% conjunction).

5.2. Basis invariance of the selected subspace

The correlated benchmark makes both the input features and their logit effects non-orthogonal. We introduce four latent block bits and assign primitive bits to blocks; each primitive copies its block bit with probability 0.8, so primitives in the same block are correlated. We also make neighboring output effects overlap by setting $c_r = 0.7q_r + 0.3q_{r+1}$, rather than assigning each feature a purely orthogonal logit direction. Two mixing matrices A produce alternate dictionaries: "mild" (random block-orthogonal 4×4 rotations on six groups of four features) and "strong" ($A = U\ \mathrm{diag}(1, 1.05, \ldots, 2.15)\, V^\top$ with random orthogonal U, V). For each dictionary, we compute the top-k right singular subspace of $\hat{T} = \hat{C}\hat{\Sigma}^{1/2}$, transport via $Q = \hat{\Sigma}'^{-1/2} A \hat{\Sigma}^{1/2}$, and compare.

Transported subspace alignment rounds to one. Across all $k \in \{1, 2, 4, 8, 12\}$, both mixings, and 8 seeds, the normalized alignment $1 - \frac{1}{2k}\|P_{V'_k} - P_{QV_k}\|_F^2$ rounds to $\mathbf{1.0000}$ at four decimal places. The empirical orthogonality of Q (i.e. $\|QQ^\top - I\|_F^2$) was below 10^{-9} in every seed. Theorem 3.3 is exact for any invertible A at the population level; our ridge-regularized finite-sample estimator is transported exactly only in the orthonormal case, because the regularizer λI does not transport under non-orthonormal A. Empirically, the effect of $\lambda = 10^{-4}$ at $N = 100{,}000$ samples is below the reported rounding in both mixings.

Named-coordinate rankings are dictionary-sensitive. In contrast to the transported whitened subspace, the set of named top-k coordinates changes substantially under reparameterization. The named-coordinate top-k overlap ranges from 0.25 to 0.50 under mild mixing and from 0.09 to 0.53 under strong mixing, depending on k. Thus, coordinate-level rankings are not stable under dictionary changes, even though the whitened subspace selected by the operator is stable.

Robustness to dictionary mismatch. The exact reparameterization $\phi' = A\phi$ tests the invariance theorem under its stated algebraic conditions. To assess robustness when the alternate dictionary is only approximately related by an invertible change of basis, we add additive Gaussian noise ($n = 8$ seeds, $k = 4$): $\phi' = A\phi + \sigma\mathcal{N}(0, I)$ with $\sigma \in \{0, 0.05, 0.1, 0.2, 0.5, 1.0\}$. The principal-angle overlap Ω_4 between the noisy-dictionary top-4 right singular subspace and the clean-ϕ' top-4 is $\{1.000, 1.000, 0.999, 0.993, 0.973, 0.962\}$, respectively. The top-$k$ whitened subspace remains stable even when the noise standard deviation is comparable to the feature scale.

5.3. Causal intervention on commonness and usefulness

The basis-invariance benchmark above showed that the whitened interaction subspace is the basis-invariant object. The causal-intervention benchmark strengthens the claim from correlation to causation by asking whether within-

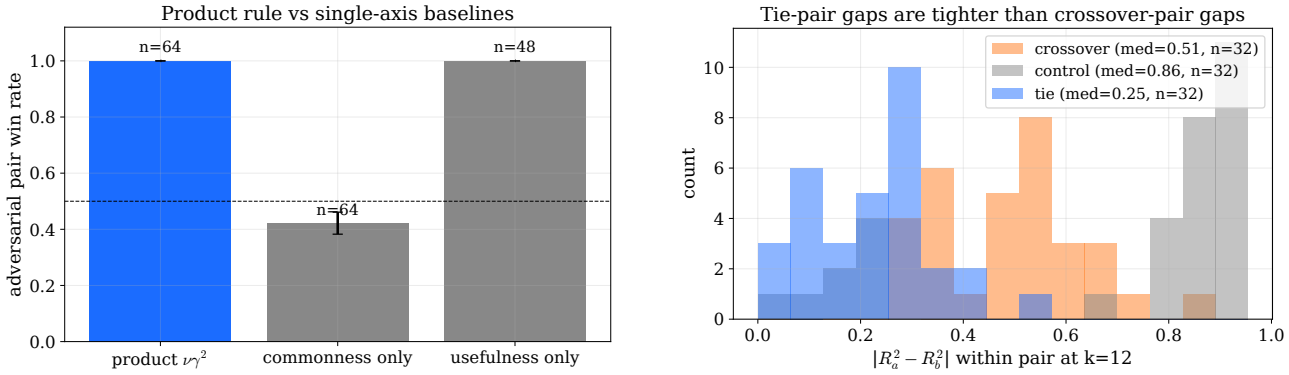

Figure 1. Orthogonal Gaussian benchmark, 8 seeds. Left: adversarial feature pairs test whether emergence is predicted by the product score $d = \nu\gamma^2$, rather than commonness ν or usefulness γ^2 alone. The product rule correctly predicts the higher-R^2 feature on all non-tie pairs, while commonness-only is inaccurate on the crossover family. Right: tie pairs, whose product scores are equal by construction, are evaluated by the within-pair readability gap $|R_a^2 - R_b^2|$ rather than ranking accuracy. Their smaller gaps show the expected no-strict-ordering behavior: equal-product features become more similarly readable than crossover or control pairs.

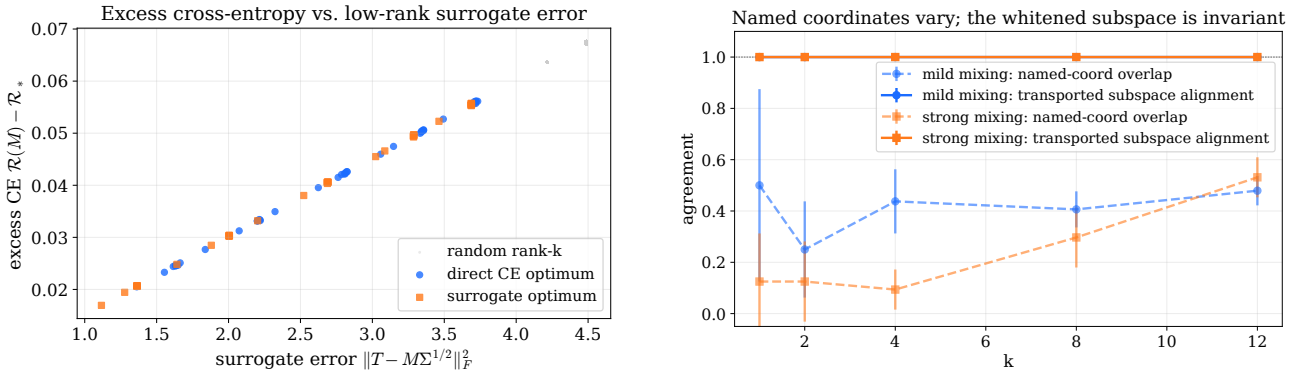

Figure 2. (Left) Excess cross-entropy vs. surrogate error pooled across (k, optimizer, seed) trained optima (○ direct CE, □ surrogate) and 100 random rank-k matrices per configuration (faint gray). The approximately linear coupling is consistent with the bounded-logit sandwich of Theorem 3.1. (Right) Basis-invariance experiment, 8 seeds: transported subspace alignment is 1.000 for both mild and strong mixings (solid lines at the top), while named-coordinate top-k overlap (dashed) varies substantially with the dictionary.

pair emergence order is *controlled by* the product score $d_r = \nu_r\gamma_r^2$. The design rules out commonness-only and usefulness-only explanations on the same feature set.

We construct eight matched feature pairs and pair each with two teacher-generator conditions A and B. Four *usefulness-flip* pairs hold (ν_a, ν_b) equal across conditions and swap (γ_a^2, γ_b^2) between conditions; four *commonness-flip* pairs hold (γ_a^2, γ_b^2) equal across conditions and swap (ν_a, ν_b) between conditions. Within each pair the product score $d_r = \nu_r\gamma_r^2$ flips between conditions. Exactly one of ν or γ^2 – the factor specific to that family – flips with it, while the other is held fixed.[1] Commonness $\nu_r = \mathbb{E}[\phi_r^2]$ is an input-distribution quantity, so a commonness flip is implemented by rescaling the primitive-feature variances between A and B while leaving the teacher's output directions c_r fixed; a usefulness flip is implemented by swapping c_r between pair members while leaving the input distribution fixed. In both cases, the two conditions share the same random output-basis $\{q_r\}$ and the same primitive seed, so the only per-coordinate difference between A and B is the single factor being flipped. For each (seed, condition) we train rank-k students at $k \in \{1, 2, 4, 8, 12\}$ against the true cross-entropy, measure probe R^2 per feature, and record the realized within-pair winner. Product-rule winners are pre-specified before training and evaluated against held-out probe results after training.

We report a paired difference-in-differences (DiD) statistic. For each pair, DiD sums the R^2 advantage of the product-predicted higher-ranked feature over the lower-ranked feature in conditions A and B. The higher-ranked feature in each condition is the feature predicted to emerge first by the product rule $d_r = \nu_r\gamma_r^2$ under that condition. This inter-

[1] Exact (ν, γ^2) values: usefulness-flip pairs use $\gamma^2 \in \{2.0, 0.2\}$ at $\nu \in \{0.3, 0.5, 0.7, 0.9\}$; commonness-flip pairs use $\gamma^2 = 1.0$ with $\nu \in \{0.3, 0.9\}, \{0.2, 0.8\}, \{0.15, 0.85\}, \{0.4, 0.7\}$.

vention is diagnostic because each single-axis ranking can succeed only on the family in which its axis is manipulated. The product rule, by contrast, makes a single prediction across both usefulness-flip and commonness-flip families.

Result. Across 8 seeds and both pair types, every pair in every condition has the product-predicted feature emerging first. At the bottleneck width $k = 8$ nearest the emergence frontier, the mean DiD statistic across pairs is

$$\mathrm{DiD}_{\text{usefulness-flip}} = +\mathbf{1.94}, \qquad \mathrm{DiD}_{\text{commonness-flip}} = +\mathbf{1.92}$$

out of a maximum of $+2.0$. At $k = 4$ the signal is smaller but unambiguous (DiD $= +1.43$ and $+0.52$, respectively), and at $k = 12$ both drop as all features saturate to $R^2 \approx 1$. In the commonness-flip family the usefulness-only rule is tied (γ^2 is held equal by construction) and indifferent on that family; in the usefulness-flip family the commonness-only rule is tied. Only the product rule flips with the observed winner in both families.

Thus, the observed DiD pattern separates the product rule from either single-axis explanation. Usefulness-only can account for the usefulness-flip family, where usefulness is the manipulated factor, but it cannot explain the commonness-flip family, where usefulness is held fixed. Conversely, commonness-only can account for the commonness-flip family but not the usefulness-flip family. The product rule is therefore the only one of the three rankings that predicts the observed winner in both conditions of both families with a single scoring function, and the only ranking that attains DiD values close to the maximum across both families simultaneously.

Figures and the complete per-pair DiD table with per-seed 95% CIs are in Appendix M.2. Per-pair DiD at $k=8$ ranges from $+1.89$ to $+1.96$ with CIs of ± 0.01–± 0.03 across 8 seeds; at $k = 4$ the range is $+0.04$ to $+1.96$, reflecting that smaller-d pairs have not fully separated at that width. Predicted winners were fixed before evaluating condition B.

5.4. Prospective subspace forecasting in a controlled transformer

We train a 4-layer decoder-only transformer on the 26-token synthetic next-token task (Appendix B) and test whether the $\hat{T}_\ell$-based whitened interaction subspace at an early forecast checkpoint forecasts the subspace that later becomes linearly readable. The central question is whether the pulled-back operator, estimated early in training, forecasts which feature subspace the transformer will later make linearly readable. Throughout this subsection, Ω_k denotes the average squared principal-angle overlap between two k-dimensional subspaces; $\Omega_k = 1$ means identical subspaces, while the random-subspace baseline is approximately k/m. We write t_0 for an early forecast training step and $t_1 = 20{,}000$ for the final training step. Per-layer operators $\hat{T}_\ell = \hat{J}_\ell \, \hat{B}_\ell \, \hat{\Sigma}^{1/2}$ are estimated by ridge, and all predictions are fixed before later-checkpoint probes are evaluated. The theory is subspace-level (Theorem 3.3), so the primary metrics are principal-angle overlaps rather than coordinate rankings.

Subspace preservation through training. At the reference configuration (p_{95}-calibrated, 4-layer transformer, layer 2, training step $t_0 = 2000$, $k = 4$, $n = 16$ seeds), the top-k right singular subspace of $\hat{T}_\ell(t_0)$ matches the top-k right singular subspace of $\hat{T}_\ell(t_1)$ at principal-angle overlap $\Omega_4^{\text{early vs late}} = \mathbf{0.74 \pm 0.07}$ (chance $k/m = 4/24 \approx 0.17$). The same hidden-state direction families $D_k(t_0) = \mathrm{orth}(\hat{B}_\ell \hat{\Sigma}^{1/2} V_k(t_0))$ overlap $D_k(t_1)$ at $\Omega_4^{\text{hidden}} = \mathbf{0.67 \pm 0.06}$, compared with hidden-PCA at 0.37 ± 0.06 and random 0.02 ± 0.00 (Appendix M.5). Preservation is above chance at every non-embedding layer and every $k \in \{1, 2, 4, 8\}$ (complete sweep in Appendix M.3).

Predicted vs empirically readable subspace. At the same reference configuration, the predicted subspace $V_4(t_1)$ matches the top-4 eigenspace of the empirical readability operator $\hat{R}_\ell = \hat{\Sigma}_{\tilde{\phi}h}(\hat{\Sigma}_{hh} + \lambda_h I)^{-1} \hat{\Sigma}_{h\tilde{\phi}}$ at $\Omega_4^{\text{pred vs read}} = \mathbf{0.51 \pm 0.07}$ (chance 0.17). The complete per-layer table is reported in Appendix M.7.

Future-only forecast. A feature-coordinate mask that zeros the already-emerged features isolates the interaction directions not yet linearly readable at forecast step t_0. Restricted to those coordinates, the early-vs-late whitened-subspace overlap is $\Omega_4^{\text{future}} = \mathbf{0.66 \pm 0.04}$ at the reference configuration. Per-seed chance baselines vary because the emerged-feature count r_0 varies across seeds; the seed-clustered gap $\Omega_4^{\text{future}} - \text{chance} = 0.170 \pm 0.057$ is positive on 13/16 seeds, rising to 0.248 ± 0.047 (16/16 seeds) under a tighter $R^2 \geq 0.9$ emergence rule (Appendix M.4).

5.5. Local-control benchmark for pulled-back directions

With the 4-layer transformer frozen, we choose a hidden-state direction δh and inject the scaled perturbation $\varepsilon \, \delta h$ at the query position of layer ℓ, where ε controls the intervention strength. We compare four steering direction families: the top singular directions of $\hat{T}_\ell$, probe-weight directions for matched feature targets, mean-difference directions, and random orthogonal baselines. The reference configuration uses $\ell = 2$, $k = 4$, $n = 16$ seeds, and $\varepsilon \in \{0.25, 0.5, 1.0, 2.0\}$. For confidence intervals, we average the 4 directions and 4 intervention strengths within each seed. Table 2 reports the three steering metrics. The singular family is best on all three metrics, and all 95% CIs for the singular family are non-overlapping with the corresponding CIs of every other family.

This pattern is not confined to the selected layer. Across the

Table 2. Local-control benchmark at the reference configuration ($\ell = 2$, $k = 4$, 16 seeds). Bold marks the best residual-stream steering family.

Metric	Sing.	Mean	Probe	Rand.
Response	$\mathbf{0.129 \pm 0.004}$	0.104 ± 0.004	0.037 ± 0.003	0.041 ± 0.002
Selectivity	$\mathbf{0.76 \pm 0.03}$	0.41 ± 0.09	0.55 ± 0.06	0.03 ± 0.06
Fidelity	$\mathbf{0.78 \pm 0.03}$	0.50 ± 0.10	0.42 ± 0.07	0.06 ± 0.05

complete $\ell \in \{1, 2, 3, 4\}$, $k \in \{1, 2, 4, 8, 12\}$ sweep ($n = 16$ seeds per cell), the singular family ranks first on fidelity at 19 of 20 configurations, with mean fidelity ratios of $\mathbf{6.57}$ relative to probe and $\mathbf{1.34}$ relative to mean-difference (Appendix M.1). The complete reference-configuration table is in Appendix L.2.

Appendix L.4 reports two supporting checks. First, a matched-total-change Pareto analysis shows that the singular family yields larger target lift at matched total perturbation. Averaged across the 20 (ℓ, k) configurations, it gives $1.25\times$ the target lift of probe and $1.75\times$ the target lift of mean-difference at matched total change. The mean-difference ratio is strictly above 1 at every configuration; the probe ratio is above 1 at 19 of 20 configurations.

Second, across the 48 (seed, layer) configurations at $k = 4$ drawn from 16 seeds, the Spearman rank correlation between the relative tail-linearity distortion $\hat{\delta}_\ell / \|\hat{J}_\ell h\|$ and the singular family's fidelity cosine is $\rho = \mathbf{-0.92}$. The seed-clustered percentile bootstrap 95% CI is $[-0.94, -0.89]$ over 2000 seed-resamples, and every replicate returns a negative correlation. The sign is consistent with the prediction that pullback fidelity degrades with tail non-linearity. Thus, the directions predicted to be preserved by cross-entropy are not merely readable: they are also effective causal control directions, producing stronger, more selective, and more locally predictable logit changes when used as steering interventions.

5.6. Prospective validation in pretrained language models

We apply the same pulled-back operator construction to Pythia models with 70M, 160M, and 410M parameters [7], accessed through the Hugging Face Transformers library [8], on a subject–verb-agreement benchmark with attractors. The benchmark uses 20 morphologically regular singular and plural noun pairs, four prepositional templates, and seven verb triples, yielding 1,344 training and 896 test examples. The lexical-plus-template hold-out requires generalization across both lexical items and syntactic frames. All operators are fit on the training split alone, and every reported metric is evaluated on the held-out split. Success therefore requires transfer across both words and templates rather than memorization of the minimal-pair set. The margin signal is $m(x) = \log p(\text{correct}) - \log p(\text{incorrect})$ at

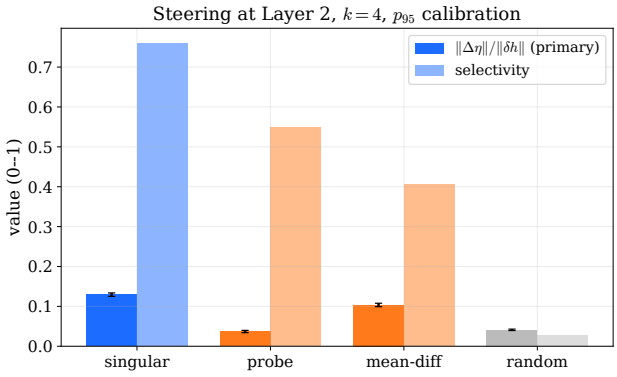

Figure 3. Reference-configuration steering results for layer 2, $k = 4$, 95th-percentile calibration on the 4-layer transformer, 16 seeds. The figure reports $\|\Delta\eta\|_2 / \|\delta h\|_2$ and selectivity for the four steering direction families. The singular family has the largest logit-change-per-perturbation and selectivity values; numerical values for all three metrics, including linearity fidelity, are in Table 2.

the last-token position, with per-example margin gradients collected at each layer. Stacking held-out margin gradients as rows of G_{te}, we report gradient-energy capture $\|G_{\text{te}} P_D\|_F^2 / \|G_{\text{te}}\|_F^2$ for a candidate subspace D, where P_D is the orthogonal projector onto D.

Prospective forecast of the task-relevant subspace. We compute the pulled-back residual-stream subspace $D_\ell^{t_0}$ at an early Pythia checkpoint $t_0 \in \{\text{step 4k}, \text{step 16k}, \text{step 48k}\}$, freeze it, and ask whether that frozen early subspace captures final-checkpoint (step 143k) margin-gradient energy on the held-out split. Across all three Pythia sizes and all three early checkpoints, the frozen early pulled-back subspace captures 1.4–14.6$\times$ as much final-checkpoint held-out gradient energy as a frozen hidden-PCA subspace computed at the same early checkpoint. The operator therefore does not merely describe current hidden-state variance; it forecasts the task-relevant subspace the model later uses. Per-configuration numbers are in Appendix L.3.2.

Same-checkpoint gradient-energy capture. At each final checkpoint and validation-selected layer (L_6, L_{12}, L_{23}) with $k = 4$ directions out of $d \in \{512, 768, 1024\}$, the pulled-back subspace captures 22–30% (range across models; per-model 95% CIs across $n = 8$ seeds ± 0.003–± 0.004) of the held-out margin-gradient energy, versus 0.4–0.8% for random and 16–26% for hidden-PCA. The gap over hidden-PCA is largest where the task is not yet saturated. At step 4,000, pulled-back captures $6.8 \pm 1.2\times$ the hidden-PCA fraction on Pythia-70M and $10.3 \pm 1.7\times$ on Pythia-160M. The gap narrows toward the final checkpoint, where the ratio is $1.1\times$ on Pythia-160M and $1.4\times$ on Pythia-410M. This is consistent with the operator describing task-relevant structure before that structure comes to dominate hidden-state variance.

Cross-model subspace forecast. The same construction transfers across model scales. The 14-dimensional SVA feature dictionary is shared across Pythia sizes, so the whitened feature-space top-4 subspaces in different models are directly comparable. For each ordered pair $(\text{small}, \text{big}) \in \{(70, 160), (70, 410), (160, 410)\}$, the small-model early-checkpoint top-4 subspace and the large-model final-checkpoint top-4 subspace overlap at $\Omega_4 = 0.66$–0.75 (chance 0.29, random-subspace baseline 0.27): the 70M step-4,000 subspace forecasts the 160M-final subspace at $\Omega_4 = 0.66$, and the 160M step-4,000 subspace forecasts the 410M-final subspace at $\Omega_4 = 0.75$. The whitened interaction subspace is therefore preserved approximately across both training time and model scale, not only within a single training run (Appendix M.6).

A structurally distinct reflexive-agreement benchmark (*himself/themselves* with an opposite-number attractor) reproduces the pattern. Ratios of pulled-back to hidden-PCA performance are $46\times$ and $4.8\times$ for Pythia-70M and Pythia-160M at the same-checkpoint final revision. The frozen-early-subspace early-to-late forecast on reflexive gives ratios of pulled-back to hidden-PCA performance of $1.5\times$–$42\times$ across six (model, t_0) configurations. The same spectral object that predicts readability and steering in controlled models therefore also identifies task-relevant directions in pretrained language models, especially before those directions dominate raw hidden-state variance. Complete tables are in Appendix L.3 and Appendix M.8.

6. Conclusion

We provide an objective-level account of why some interaction features become linearly readable before others under next-token cross-entropy. Bounded-logit cross-entropy under a width-k bottleneck is shown to be uniformly comparable to the low-rank approximation error of a single spectral object, the whitened interaction-to-logit operator $T = C\Sigma^{1/2}$, so that the bottleneck preserves its top-k right singular subspace. Pre-linearization, usually treated as a qualitative observation, becomes an explicit SVD problem.

Equivalently, cross-entropy under a bottleneck selects whitened interaction directions with high logit-wise predictive leverage; in orthogonal coordinates, this selection reduces to the commonness-times-usefulness score $d_r = \nu_r \|c_r\|_2^2$. This makes explicit how a rare but highly predictive feature can outrank a common but weakly predictive one. In the general case the invariant object is not a list of named features but a basis-invariant whitened interaction subspace, preserved under any invertible reparameterization of the dictionary. At middle layers the same geometry reappears as the pulled-back operator $T_\ell = J_\ell B \Sigma^{1/2}$, whose top singular directions are first-order optimal for local steering per unit whitened-feature perturbation. Readability and local control are therefore governed by a single object.

The empirical program tests the same spectral predictions across settings of increasing realism and couples readability to local control throughout. On the adversarial-pair benchmark, the product rule selects the higher-readability feature for every Gaussian non-tie pair and remains the strongest ranking under pairwise-conjunction features; commonness-only is inaccurate on the crossover family in both variants. The basis-invariance experiment shows transported subspace alignment numerically equal to one even as named-coordinate overlap varies substantially, corroborating that the subspace, not the coordinate list, is canonical. A matched-pair intervention strengthens this from correlation to causation: flipping either commonness or usefulness flips the observed emergence winner exactly as the product score predicts.

At middle layers, the controlled-transformer experiments show that the pulled-back operator forecasts later-checkpoint readable subspaces, outperforming hidden-PCA and random baselines, and yields more selective steering directions than probe, mean-difference, and random baselines. On pretrained Pythia, the same operator forecasts the task-relevant subspace on held-out subject–verb and reflexive agreement minimal pairs. It outperforms hidden-PCA and random baselines on held-out margin-gradient energy before the task-relevant structure comes to dominate hidden-state variance.

Across these settings, the same spectral object predicts feature selection, subspace preservation, linear readability, and steering. This supports the conclusion that pre-linearization is not merely a convenient description of trained representations, but an objective-level implicit bias of cross-entropy under bottlenecks.

More broadly, our results also suggest a design principle for objectives aimed at interpretability or steerability. We show that next-token cross-entropy acts as an implicit pre-linearization pressure under bottlenecks. This suggests that continued-training or post-training objectives may benefit from retaining a cross-entropy or KL-to-base term, rather than optimizing only a downstream behavioral objective. Such a term could encourage predictive structure to remain concentrated in low-dimensional, linearly accessible subspaces. We view this as a design hypothesis motivated by the theory, and as a direction for future work.

Impact Statement

This paper studies why predictive structure in neural language models becomes linearly readable under representation bottlenecks. The work is primarily theoretical and diagnostic, and its intended positive impact is to improve scientific understanding of representation formation, inter-

pretability, and controllability in neural networks. We do not anticipate any direct societal consequences of this work.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop Track*, 2017.

[2] John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138, 2019.

[3] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[4] Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models. In *International Conference on Learning Representations*, 2025.

[5] Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21879–21911, 2024.

[6] Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. In *Conference on Language Modeling*, 2024.

[7] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430, 2023.

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, 2020.

[9] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, volume 14, pages 617–624, 2001.

[10] Shakir Mohamed, Katherine A. Heller, and Zoubin Ghahramani. Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems*, volume 21, pages 1089–1096, 2008.

[11] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *Foundations and Trends in Machine Learning*, 9(1):1–118, 2016.

[12] Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

[13] T. W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22(3): 327–351, 1951.

[14] Gregory C. Reinsel and Raja P. Velu. *Multivariate Reduced-Rank Regression: Theory and Applications*, volume 136 of *Lecture Notes in Statistics*. Springer, 1998.

[15] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018.

[16] Sekitoshi Kanai, Yasuhiro Fujiwara, Yuki Yamanaka, and Shuichi Adachi. Sigsoftmax: Reanalysis of the softmax bottleneck. In *Advances in Neural Information Processing Systems*, volume 31, pages 286–296, 2018.

[17] Octavian Ganea, Sylvain Gelly, Gary Becigneul, and Aliaksei Severyn. Breaking the softmax bottleneck via learnable monotonic pointwise non-linearities. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2073–2082. PMLR, 2019.

[18] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[19] Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960.

[20] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[21] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[22] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27, pages 2177–2185, 2014.

[23] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Chris Olah. Toy models of superposition. Transformer Circuits Thread, 2022.

[24] Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

[25] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

[26] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

[27] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. Transformer Circuits Thread, 2024.

[28] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021.

[29] Nelson Elhage, Robert Lasenby, and Christopher Olah. Privileged bases in the transformer residual stream. Transformer Circuits Thread, 2023.

[30] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. arXiv preprint arXiv:2308.10248, 2023.

[31] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. arXiv preprint arXiv:2310.01405, 2023.

[32] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023.

[33] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

[34] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7237–7256, 2020.

[35] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

[36] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

A. Extended related work

Relation to classical low-rank objectives. Classical low-rank frameworks approach related questions from another direction. Exponential-family PCA [9, 10], generalized low-rank models [11], reduced-rank regression [12–14], and softmax-bottleneck language modeling [15–17] provide Bregman-loss and low-rank machinery. The objective reduction in this paper is related to these traditions, but differs in three respects. First, the rank constraint is imposed on an interaction dictionary. Second, the selected object is a whitened subspace rather than a coordinate list. Third, the same operator also induces a middle-layer local-control prediction.

Bregman low-rank approximation for cross-entropy. Our bounded-logit theorem (Theorem 3.1) is a uniform sandwich that reduces cross-entropy to a Frobenius low-rank surrogate on $T = C\Sigma^{1/2}$; on the Frobenius side Eckart–Young–Mirsky [18, 19] then selects the top singular subspace. It is classical that exponential families induce Bregman divergences and that PCA-style low-rank approximation extends from squared loss to these divergences. Collins, Dasgupta, and Schapire's *generalization of PCA to the exponential family* [9] is the closest predecessor: they define low-rank approximations that minimize the Bregman divergence associated with the natural-parameter form of an exponential family, which for the categorical likelihood is cross-entropy. Udell et al.'s generalized low-rank models [11] extend this to arbitrary data types and loss functions with convex penalties, and Banerjee et al. [20] develop the general Bregman-divergence theory we implicitly use. The present paper specializes to softmax cross-entropy in the next-token setting, makes the sandwich bounds quantitative in terms of the bounded-logit curvature constants m_B, M_B, and connects the resulting spectral object to interpretability-relevant quantities (readability, subspace preservation, steering).

Reduced-rank regression and CCA. Once the problem is phrased as a rank-k linear map from a feature dictionary to logits, it belongs to the classical family of reduced-rank regression (Izenman [12]) and canonical-correlation-style methods (Hotelling's CCA [21]). Izenman develops reduced-rank regression for the multivariate linear model and relates it to principal components and canonical variates. The pulled-back operator differs from classical reduced-rank regression in that the outputs are logits composed with the downstream Jacobian, so the Jacobian J_ℓ is part of the object being fit. This dependence on J_ℓ is what connects the spectral object to local control.

Softmax bottleneck and language modeling as matrix factorization. Yang et al. [15] frame language modeling as a matrix-factorization problem and argue that softmax-based LMs are rank-limited. Our bottleneck is different (the rank constraint is on the interaction-to-logit map, not on the context-to-logit matrix), but the shared ingredient is that the softmax output distribution induces rank constraints on the pre-softmax structure it can represent. Our T is a fixed-dictionary operator whose top singular directions are the ones the bottleneck preserves. Levy and Goldberg [22] show that Skip-Gram with negative sampling implicitly factorizes a shifted PMI matrix; more generally, their paper is the standard citation for the view that NLP objectives induce weighted matrix-factorization problems. The present paper studies an analogous implicit-factorization structure for bounded-logit cross-entropy with a fixed interaction dictionary, where the singular directions have a direct interpretation as selected interaction subspaces.

Next-token prediction and linear representations. Jiang et al. [5] study how latent-variable structure can give rise to linear concept representations under next-token prediction with softmax cross-entropy and gradient descent: if the data-generating process has binary latent concepts that are independent in a precise sense, the softmax-cross-entropy gradient pushes their unembeddings to be linearly related. In contrast, our analysis starts from an interaction dictionary and characterizes which subspace of that dictionary is selected under a bottleneck, via the operator $T = C\Sigma^{1/2}$ and its middle-layer pullback. Zhao et al. [6] relate next-token prediction to rank-constrained and nuclear-norm-regularized optimization in the logit domain and to a sparse-plus-low-rank factorization. Our paper differs in that we start from the interaction dictionary and a width-k bottleneck, and we tease out the whitened pulled-back singular subspace as a readable object, tested with a readability probe and a steering benchmark. Merullo et al. [4] link pretraining frequency to linear representability; our commonness-times-usefulness law strictly generalizes frequency-based accounts by giving predictive leverage an explicit role, and the causal-intervention experiment separates the product rule from commonness-only and leverage-only alternatives in the matched intervention design. In fact, our experiments suggest that logit-wise predictive leverage is at least as important, and often more discriminative, when the two factors are experimentally separated (Sections 5.1 and 5.3). Cross-entropy does not select frequent features per se; it selects directions with large commonness-times-usefulness score.

Superposition, feature importance, and sparse-feature dictionaries. The product rule $d_r = \nu_r \|c_r\|_2^2$ is related to capacity-allocation accounts in superposition models: features compete for limited representational capacity, and both frequency and predictive importance matter. Elhage et al.'s *toy models of superposition* [23] describe this kind of competition under limited dimensionality, with feature importance and sparsity trading off against each other. The quantities differ: ν_r is a second moment and $\|c_r\|_2^2$ is predictive leverage in the fixed dictionary, neither of which is a sparsity count or an external importance score. The structural parallel is that a capacity-constrained objective allocates directions to features with large frequency-leverage product. Huben et al. [24] and subsequent sparse-autoencoder work (Bricken et al. [25], Gemma Scope [26], Templeton et al. [27]) supply empirical dictionaries of interpretable directions. Our contribution is complementary: we identify the spectral law by which cross-entropy allocates capacity among directions of an interaction dictionary. The law is obtained directly from the cross-entropy objective through a whitened low-rank reduction, yielding a basis-invariant singular subspace rather than a coordinate-level heuristic selection rule suggested in prior works.

Basis invariance and "no privileged basis". Our Theorem 3.3, that invertible reparameterizations of the dictionary preserve the selected whitened subspace, is mathematically a whitening-equivariance fact. Its interpretability significance is that, under the residual-stream linear symmetries documented by Elhage et al. [28], named coordinates of a dictionary are not canonical objects; the basis-invariant whitened subspace is. Separately, empirical work on privileged bases under realistic training [29] shows that optimizer and parameterization details can induce privileged directions absent from the idealized linear theory. This empirical observation is compatible with our basis-invariance claim. Cross-entropy selects a basis-invariant whitened subspace, while training dynamics and parameterization can shape the coordinate basis in which that subspace is most readable.

Steering, representation engineering, and concept erasure. Our steering experiments ask whether the top singular directions of T_ℓ are also the best directions for local intervention under the appropriate perturbation norm. The broader line of work on activation-level control includes activation engineering (Turner et al. [30]), representation engineering (Zou et al. [31]), inference-time intervention (Li et al. [32]), and unsupervised latent-knowledge discovery (Burns et al. [33]). These works identify activation-space directions or representations from supervised or unsupervised signals; the steering methods then intervene along such directions. Our construction is theoretically motivated: it uses the same pulled-back operator that governs readability. We evaluate the resulting directions against probe, mean-difference, and random baselines in controlled and pretrained settings. On the erasure side, Ravfogel et al.'s INLP [34] and Belrose et al.'s LEACE [35] identify linear concept subspaces and remove them. This is dual to our steering question. The top singular directions identify subspaces favored by the induced objective for encoding the target interaction, making them natural candidates for both selective amplification and selective removal. We test amplification here and leave erasure to future work.

Dynamics. Appendix I analyzes aligned balanced equal-amplitude gradient flow for the factorized surrogate, showing that larger singular modes reach fixed amplitude thresholds earlier. This follows the deep-linear-network tradition of Saxe et al. [36]. Appendix I uses the factorized surrogate only to motivate mode-wise emergence order; it is not intended as a theorem about full end-to-end SGD.

Position of this work. This paper identifies a single spectral object that the next-token cross-entropy objective selects under a realistic bottleneck, together with the empirical consequences of that identification. Concretely, bounded-logit next-token cross-entropy, in the setting of an interaction dictionary passed through a width-k representation bottleneck, is uniformly sandwiched by the low-rank approximation error of the pulled-back operator $T_\ell = J_\ell B \Sigma^{1/2}$. The top-$k$ right singular subspace is basis-invariant (Theorem 3.3) and first-order optimal for local control per unit whitened-feature perturbation (Theorem K.10). The hidden-state Euclidean-norm analogue is a generalized-eigenvalue problem (Proposition K.11). In orthogonal coordinates this subspace is ranked by the commonness-times-usefulness product $d_r = \nu_r \|c_r\|_2^2$ (Corollary 3.4). The matched-pair causal intervention then separates this product score from commonness-only and usefulness-only explanations, and from prior frequency-based accounts of linear representability. Prior work identifies frequency as a linearization selection pressure, but our experiments suggest that logit-wise predictive leverage is at least as important, and often more discriminative, when the two factors are experimentally separated (Sections 5.1 and 5.3). Cross-entropy does not select frequent features per se; it selects directions with large commonness-times-usefulness score.

The accompanying empirical program spans synthetic teacher–student settings, a controlled transformer trained on a known dictionary, and pretrained Pythia models on subject–verb agreement. It tests four predictions jointly: crossover, basis invariance, subspace preservation across training, and selective steering. To our knowledge, no prior work ties all

four together through a single spectral object selected by the training objective. The causal-intervention and readability experiments further support the prediction at the level of individual interaction features. Together, these results connect crossover behavior, basis invariance, training-time subspace preservation, and selective steering to one objective-selected spectral object.

B. Exact implementation protocol

This appendix specifies the datasets, model configurations, optimization settings, and evaluation protocols used in the experiments.

B.1. Data splits and random seeds

All experiments use 8 seeds unless stated otherwise. The synthetic teacher–student experiments use $200{,}000$ training examples, $20{,}000$ validation examples, and $50{,}000$ test examples per seed. The controlled transformer trains on freshly generated batches for $20{,}000$ steps with batch size 512, exposing the optimizer to roughly 10^7 training tokens. Evaluation uses fixed validation and test sets of $20{,}000$ sequences each per seed, sampled once and held constant across all checkpoints within a seed.

B.2. Orthogonal and conjunction synthetic benchmark

Orthogonal Gaussian benchmark. The fixed orthogonal Gaussian feature set consists of 24 independent Gaussian features in three groups of eight:

- **Tie pairs (8 features, 4 pairs).** Each pair uses $(\nu, \gamma^2) = (0.64, 0.25)$ and $(0.16, 1.0)$.
- **Crossover pairs (8 features, 4 pairs).** Each pair uses $(\nu, \gamma^2) = (0.64, 0.125)$ and $(0.16, 1.0)$.
- **Control pairs (8 features, 4 pairs).** Two same-usefulness pairs and two same-commonness pairs, chosen so that both simple baselines are sometimes right and sometimes wrong.

Twenty-four orthonormal $q_r \in G \subset \mathbb{R}^{33}$ are drawn by QR decomposition on a Gaussian matrix projected into the zero-sum gauge. The teacher logits are

$$\eta_*(x) = \sum_{r=1}^{24} \gamma_r \phi_r(x) q_r,$$

rescaled by a single global constant so that the maximum observed $\|\eta_*(x)\|_\infty$ on the combined training and validation set is at most 2.5.

Conjunction replication. The conjunction variant uses 12 primitive Bernoulli variables with

$$\begin{aligned}(p_1, \ldots, p_{12}) = (&0.05, 0.05, 0.1, 0.1, 0.2, 0.2,\\ &0.35, 0.35, 0.5, 0.5, 0.2, 0.35),\end{aligned}$$

and 24 pairwise conjunction features $\phi_r(x) = (x_i - p_i)(x_j - p_j)$ on selected primitive pairs. These features reuse the usefulness weights γ_r of the Gaussian benchmark; realized variances $\hat{\nu}_r$ are measured on the training set and the tie and crossover comparisons use the realized scores $\hat{d}_r$.

Optimization defaults. For direct low-rank cross-entropy optimization, $M = UP$ with $U \in \mathbb{R}^{n\times k}$ and $P \in \mathbb{R}^{k\times m}$; entries of U and P are initialized IID from $\mathcal{N}(0, 10^{-3})$. Training uses AdamW, learning rate 10^{-2}, batch size $4{,}096$, weight decay 10^{-4}, and $5{,}000$ steps. The surrogate uses the same parameterization and optimizer with learning rate 3×10^{-3} and $3{,}000$ steps.

Random rank-k baselines. To visualize risk–surrogate coupling away from optimum, we sample 100 random pairs (U, P) per seed and width at the same initialization scale and compute both excess risk and surrogate error on the test set, without optimizing them.

B.3. Correlated and basis-mixed synthetic benchmark

The correlated benchmark reuses the conjunction-feature setting with four latent block bits $z_1, \ldots, z_4 \sim \mathrm{Bernoulli}(1/2)$. Three primitives are assigned to each block. Each primitive equals its block bit with probability 0.8 and is flipped otherwise. For basis invariance we use two mixing families:

- **Mild mixing:** random block-orthogonal 4×4 rotations applied independently to six groups of four features;
- **Strong mixing:** $A = U\,\mathrm{Diag}(1, 1.05, \ldots, 2.15)V^\top$ with random orthogonal U, V.

For each dictionary, we compute the top-k right singular subspace of $\hat{T}$ and transport the resulting basis back to the original feature space via the whitened coordinate relation in Theorem 3.3. Reported quantities are named top-k overlap between dictionaries, principal angles between transported singular subspaces, and the projector distance from each predicted subspace to the empirical readable subspace.

B.4. Controlled transformer with a known interaction dictionary

Training data generation. Each sequence is generated as follows: the 12 primitive bits are sampled; they are written into token positions 1–12 using dedicated tokens for bit value 0 and bit value 1 at each position; positions 13–24 are filled with IID noise tokens from a 16-token alphabet; a fixed query token is placed at position 25; the label token at position 26 is sampled from the teacher softmax. The loss is applied only to the position-25 prediction of token 26.

Model checkpoints. We evaluate checkpoints every 200 training steps and use validation-set residual streams, logits, and dictionary features for probe estimation. At each checkpoint the query-token residual stream at all four layers, the logits, and the exact dictionary feature matrix on the fixed validation set are retained.

Pre-specified predictions. At each forecast checkpoint $t_0 \in \{400, 1000, 2000\}$, we generate predicted feature ranks, top-k subspace bases, first-emergence layers, and coarse emergence-time buckets. These predictions are fixed before later-checkpoint probe results are evaluated.

Probe suite. For each checkpoint and each layer, scalar ridge probes are fit from the residual stream to every named feature; probes are also fit to the top singular coordinates of $\hat{T}_\ell$; $\hat{R}_\ell$ and its top-k eigenspace are computed; and we compute R^2 values and projector matrices for each checkpoint. The ridge regularization schedule is held fixed across checkpoints, so changes over time are not driven by probe hyperparameters.

B.5. Frozen-tail steering benchmark

Layer choice. The reported layer sweep is $\ell \in \{1, 2, 3, 4\}$ in the 4-layer model; layer 2 is the default reporting layer and the others are robustness checks.

Intervention directions. For each layer and each target rank k:

- **Singular directions:** top right singular vectors of $\hat{T}_\ell$ mapped into hidden-state space through the estimated bottleneck map;
- **Probe directions:** normalized ridge probe weight vectors;
- **Mean-difference directions:** mean activation on feature-on examples minus mean activation on matched feature-off examples;
- **Random controls:** matched-norm random orthogonal directions.

Steering strengths. For each direction δh we test $\epsilon \in \{0.25, 0.5, 1.0, 2.0\}$ and measure effects on 2,000 examples drawn from the eval half of a 50/50 split of the 20,000-example fixed validation set. The pulled-back operator $\hat{T}_\ell$, the probe weights, the mean-difference directions, and the top-k feature selection are all derived from the disjoint 10,000-example fit half; the same eval examples are then used for every method to reduce noise.

Target and off-target scores. For each intervention we record:

1. $\Delta_{\text{target}} = \langle \eta^{\text{steered}} - \eta^{\text{base}}, u_i \rangle$ for the intended target left singular direction or teacher logit direction;
2. $\Delta_{\text{total}} = \|\eta^{\text{steered}} - \eta^{\text{base}}\|_2$;
3. selectivity $= \Delta_{\text{target}}/(\Delta_{\text{total}} + 10^{-8})$;
4. KL divergence between the original and steered predictive distributions;
5. cosine similarity between the actual logit change and the linear prediction $\hat{J}_\ell \delta h$.

Best-layer selection protocol. The layer reported in the main text is the one that maximizes the *singular family's* fidelity cosine at $k = 4$ on the validation split. This selection rule favors the layer at which the singular family is strongest. We therefore also report the full (ℓ, k) sweep in Appendix M.1. The sweep shows that the singular family exceeds probe and random baselines at every one of the 20 (ℓ, k) cells, and exceeds mean-difference at 19 of 20, so the reported configuration is representative. The per-seed CIs at the reference configuration are computed over seeds only; the layer and k are fixed.

C. Setup and notation

This appendix collects notation used in the theoretical appendices.

C.1. Probability space and logits

Let (X, μ) be a probability space of contexts. Let $n \geq 2$ be the vocabulary size. Because the softmax map is invariant to adding the same constant to every logit, we work in the standard zero-sum gauge

$$G := \{z \in \mathbb{R}^n : \mathbf{1}^\top z = 0\}.$$

We equip G with the Euclidean inner product inherited from $\mathbb{R}^n$.

The teacher next-token distribution is represented by a measurable map

$$\eta_* : X \to G, \qquad p_*(\cdot \mid x) := \mathrm{softmax}(\eta_*(x)).$$

We assume $\eta_* \in L_2(\mu; G)$.

Define the log-sum-exp potential

$$A(z) := \log\Big(\sum_{i=1}^{n} e^{z_i}\Big), \qquad z \in \mathbb{R}^n.$$

Its gradient is the softmax probability vector

$$\nabla A(z) = \mathrm{softmax}(z) =: p(z),$$

and its Hessian is

$$H(z) := \nabla^2 A(z) = \mathrm{Diag}(p(z)) - p(z)p(z)^\top.$$

C.2. Interaction dictionary

Fix an interaction dictionary

$$\phi = (\phi_1, \ldots, \phi_m) : X \to \mathbb{R}^m.$$

We assume throughout that each $\phi_r \in L_2(\mu)$ and that the dictionary second-moment matrix

$$\Sigma := \mathbb{E}[\phi(X)\phi(X)^\top]$$

is positive definite.

Remark C.1. The assumption $\Sigma \succ 0$ excludes exact linear dependence inside the chosen interaction dictionary. This is a notational convenience rather than a conceptual restriction: every theorem below extends to singular Σ by replacing $\Sigma^{-1/2}$ with the Moore-Penrose pseudoinverse and restricting all feature-space statements to $\mathrm{range}(\Sigma)$.

Define the whitened dictionary

$$\widetilde{\phi}(x) := \Sigma^{-1/2}\phi(x).$$

Then

$$\mathbb{E}[\widetilde{\phi}(X)\widetilde{\phi}(X)^\top] = I_m.$$

C.3. Dictionary-generated logit family

Consider the finite-dimensional linear subspace of $L_2(\mu; G)$ given by

$$\mathcal{F} := \{M\phi(\cdot) : M \in \mathbb{R}^{n\times m},\ \mathbf{1}^\top M = 0\}.$$

Because $\mathcal{F}$ is finite-dimensional, it is a closed linear subspace of the Hilbert space $L_2(\mu; G)$. Therefore the orthogonal projection of η_* onto $\mathcal{F}$ exists. We denote it by

$$P_{\mathcal{F}}\eta_* = C\phi, \qquad \mathbf{1}^\top C = 0,$$

and we define the residual

$$r := \eta_* - C\phi.$$

By construction, $r \perp \mathcal{F}$ in $L_2(\mu; G)$.

Finally, define the whitened interaction-to-logit operator

$$T := C\Sigma^{1/2} \in \mathbb{R}^{n\times m}.$$

Then

$$C\phi(x) = T\widetilde{\phi}(x).$$

C.4. Width-k students

A width-k dictionary bottleneck is any student logit map of the form

$$\widehat{\eta}(x) = Uh(x), \qquad h(x) \in \mathbb{R}^k, \qquad \widehat{\eta}(x) \in G,$$

where h itself is a linear function of the dictionary features. Equivalently, it is any map of the form

$$\widehat{\eta}_M(x) = M\phi(x), \qquad M \in \mathbb{R}^{n\times m}, \quad \mathbf{1}^\top M = 0, \quad \operatorname{rank}(M) \le k.$$

Indeed, $\operatorname{rank}(M) \le k$ if and only if one can factor $M = UP$ with $U \in \mathbb{R}^{n\times k}$ and $P \in \mathbb{R}^{k\times m}$, in which case

$$\widehat{\eta}_M(x) = U(P\phi(x))$$

uses the width-k representation $h(x) = P\phi(x)$.

The population next-token cross-entropy risk of M is

$$\mathcal{R}(M) := \mathbb{E}_{X,Y\sim p_*}\big[-\log \operatorname{softmax}(M\phi(X))_Y\big].$$

The Bayes risk is

$$\mathcal{R}_* := \mathbb{E}_{X,Y\sim p_*}\big[-\log p_*(Y \mid X)\big].$$

D. Preliminary lemmas

We collect the representation and softmax identities used to reduce cross-entropy to the matrix surrogate.

D.1. Representations and readable subspaces

Definition D.1 (Readable scalar concept space)**.** Let $h : X \to \mathbb{R}^k$ be any representation. Its linearly readable scalar concept space is

$$\mathcal{S}(h) := \{a^\top h(\cdot) : a \in \mathbb{R}^k\} \subset L_2(\mu).$$

Lemma D.2 (Representation lemma)**.** *Let $S \subset L_2(\mu)$ be a k-dimensional linear subspace. Then there exists a representation $h : X \to \mathbb{R}^k$ such that*

$$\mathcal{S}(h) = S.$$

In fact, if $g_1, \dots, g_k$ is any basis of S, then the choice

$$h(x) := (g_1(x), \dots, g_k(x))$$

has this property.

Proof. Fix a basis $g_1, \dots, g_k$ of S and define h as above.

For $\mathcal{S}(h) \subseteq S$. Take any element of $\mathcal{S}(h)$. By definition it has the form

$$a^\top h(\cdot) = \sum_{j=1}^{k} a_j g_j(\cdot)$$

for some vector $a = (a_1, \dots, a_k) \in \mathbb{R}^k$. Since S is a linear subspace containing each basis vector g_j, it also contains every linear combination of the g_j. Therefore $a^\top h(\cdot) \in S$. This proves $\mathcal{S}(h) \subseteq S$.

For $S \subseteq \mathcal{S}(h)$. Take any $f \in S$. Because $g_1, \dots, g_k$ is a basis of S, there exist unique scalars $a_1, \dots, a_k$ such that

$$f = \sum_{j=1}^{k} a_j g_j.$$

Now define $a = (a_1, \dots, a_k) \in \mathbb{R}^k$. Then for every $x \in X$ we have

$$a^\top h(x) = \sum_{j=1}^{k} a_j g_j(x) = f(x).$$

Hence $f = a^\top h(\cdot) \in \mathcal{S}(h)$. This proves $S \subseteq \mathcal{S}(h)$.

Since both inclusions hold, we conclude that $\mathcal{S}(h) = S$. □

D.2. Cross-entropy as Bregman divergence

Lemma D.3 (KL/Bregman identity for softmax)**.** *For any $u, v \in G$, let $p_u = \mathrm{softmax}(u)$ and $p_v = \mathrm{softmax}(v)$. Then*

$$\mathrm{KL}(p_u \| p_v) = D_A(v, u),$$

where

$$D_A(v, u) := A(v) - A(u) - \langle \nabla A(u), v - u \rangle$$

is the Bregman divergence of A.

Proof. By definition of the softmax distribution,

$$(p_u)_i = \frac{e^{u_i}}{\sum_j e^{u_j}} = e^{u_i - A(u)}, \qquad (p_v)_i = \frac{e^{v_i}}{\sum_j e^{v_j}} = e^{v_i - A(v)}.$$

Therefore

$$\log \frac{(p_u)_i}{(p_v)_i} = u_i - A(u) - v_i + A(v).$$

Multiply by $(p_u)_i$ and sum over i:

$$\begin{aligned}\mathrm{KL}(p_u\|p_v) &= \sum_{i=1}^{n}(p_u)_i \log \frac{(p_u)_i}{(p_v)_i} \\ &= \sum_{i=1}^{n}(p_u)_i\big(u_i - A(u) - v_i + A(v)\big) \\ &= \sum_{i=1}^{n}(p_u)_i u_i - \sum_{i=1}^{n}(p_u)_i v_i \\ &\quad - A(u)\sum_{i=1}^{n}(p_u)_i + A(v)\sum_{i=1}^{n}(p_u)_i.\end{aligned}$$

Because p_u is a probability vector, its coordinates sum to 1. So the last expression becomes

$$\mathrm{KL}(p_u\|p_v) = \langle p_u, u\rangle - \langle p_u, v\rangle - A(u) + A(v).$$

Since $\nabla A(u) = p_u$, we may rewrite this as

$$\mathrm{KL}(p_u\|p_v) = A(v) - A(u) - \langle \nabla A(u), v-u\rangle = D_A(v,u).$$

That is exactly the desired identity. □

Lemma D.4 (Integral Hessian representation of Bregman divergence)**.** *Let $A : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Then for every $u, v \in \mathbb{R}^n$,*

$$D_A(v,u) = \int_0^1 (1-t)\,(v-u)^\top \nabla^2 A\big(u + t(v-u)\big)(v-u)\ \mathrm{d}t.$$

Applied to log-sum-exp, this becomes

$$D_A(v,u) = \int_0^1 (1-t)\,(v-u)^\top H\big(u + t(v-u)\big)(v-u)\ \mathrm{d}t.$$

Proof. Fix $u, v \in \mathbb{R}^n$ and define the one-variable function

$$g(t) := A\big(u + t(v-u)\big), \qquad t \in [0,1].$$

Because A is twice continuously differentiable, so is g. By the chain rule,

$$g'(t) = \big\langle \nabla A\big(u + t(v-u)\big), v-u\big\rangle$$

and

$$g''(t) = (v-u)^\top \nabla^2 A\big(u + t(v-u)\big)(v-u).$$

Now apply the one-dimensional integral form of Taylor's theorem at $t = 0$ evaluated at $t = 1$:

$$g(1) = g(0) + g'(0) + \int_0^1 (1-t)g''(t)\ \mathrm{d}t.$$

Substituting the definitions of $g(1)$, $g(0)$, and $g'(0)$ gives

$$\begin{aligned}A(v) &= A(u) + \langle \nabla A(u), v-u\rangle \\ &\quad + \int_0^1 (1-t)(v-u)^\top \nabla^2 A\big(u + t(v-u)\big)(v-u)\ \mathrm{d}t.\end{aligned}$$

Move the first two terms on the right-hand side to the left. The result is

$$D_A(v,u) = \int_0^1 (1-t)(v-u)^\top \nabla^2 A\big(u + t(v-u)\big)(v-u)\ \mathrm{d}t.$$

This is the claimed identity. □

D.3. Explicit bounded-logit constants

For $B > 0$ define the bounded zero-sum logit box

$$K_B := \{z \in G : \|z\|_\infty \le B\}.$$

Because K_B is compact and the map $z \mapsto H(z)$ is continuous, the extremal eigenvalues of $H(z)|_G$ on K_B are finite and attained. Define

$$\begin{aligned} m_B &:= \inf_{z \in K_B} \lambda_{\min}(H(z)|_G), \\ M_B &:= \sup_{z \in K_B} \lambda_{\max}(H(z)|_G), \\ \kappa_B &:= \frac{M_B}{m_B}. \end{aligned}$$

The next proposition gives explicit usable bounds for m_B and M_B.

Proposition D.5 (Explicit Hessian bounds on a bounded logit box)**.** *Let $z \in K_B$ and let $u \in G$. Then*

$$\frac{1}{1 + (n-1)e^{2B}} \, \|u\|_2^2 \le u^\top H(z) u \le \|u\|_2^2 .$$

Consequently,

$$m_B \ge \frac{1}{1 + (n-1)e^{2B}}, \qquad M_B \le 1, \qquad \kappa_B \le 1 + (n-1)e^{2B}.$$

Proof. First note the variance identity for $p = \mathrm{softmax}(z)$:

$$\begin{aligned} u^\top H(z) u &= u^\top \big(\mathrm{Diag}(p) - pp^\top \big) u \\ &= \sum_{i=1}^n p_i u_i^2 - \Big(\sum_{i=1}^n p_i u_i \Big)^2. \end{aligned}$$

Define the p-weighted mean

$$\bar{u}_p := \sum_{i=1}^n p_i u_i.$$

Then

$$\begin{aligned} \sum_{i=1}^n p_i (u_i - \bar{u}_p)^2 &= \sum_{i=1}^n p_i u_i^2 - 2\bar{u}_p \sum_{i=1}^n p_i u_i + \bar{u}_p^2 \sum_{i=1}^n p_i \\ &= \sum_{i=1}^n p_i u_i^2 - 2\bar{u}_p^2 + \bar{u}_p^2 \\ &= \sum_{i=1}^n p_i u_i^2 - \bar{u}_p^2 \\ &= u^\top H(z) u. \end{aligned}$$

Thus

$$u^\top H(z) u = \sum_{i=1}^n p_i (u_i - \bar{u}_p)^2.$$

This also equals

$$\min_{c \in \mathbb{R}} \sum_{i=1}^n p_i (u_i - c)^2,$$

because the weighted least-squares minimizer is precisely $c = \bar{u}_p$.

We now derive the lower bound. Let

$$p_{\min}(z) := \min_i p_i.$$

For every $c \in \mathbb{R}$,

$$\sum_{i=1}^{n} p_i(u_i - c)^2 \geq p_{\min}(z) \sum_{i=1}^{n} (u_i - c)^2.$$

Take the minimum over c on both sides. Since the left-hand side equals $u^\top H(z)u$, we obtain

$$u^\top H(z)u \geq p_{\min}(z) \min_{c\in\mathbb{R}} \sum_{i=1}^{n} (u_i - c)^2.$$

The minimizer of the unweighted quadratic on the right-hand side is the ordinary mean

$$\bar{u} := \frac{1}{n} \sum_{i=1}^{n} u_i.$$

Because $u \in G$, we have $\sum_i u_i = 0$, hence $\bar{u} = 0$. Therefore

$$\min_{c\in\mathbb{R}} \sum_{i=1}^{n} (u_i - c)^2 = \sum_{i=1}^{n} u_i^2 = \|u\|_2^2 .$$

So

$$u^\top H(z)u \geq p_{\min}(z) \|u\|_2^2 .$$

It remains to bound $p_{\min}(z)$ from below using the box constraint. Since $z \in K_B$, every coordinate of z lies in $[-B, B]$. Therefore, for each i,

$$p_i(z) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \geq \frac{e^{-B}}{e^{-B} + (n-1)e^{B}} = \frac{1}{1 + (n-1)e^{2B}}.$$

Combining the last two displays yields the announced lower bound.

For the upper bound, use again the identity

$$u^\top H(z)u = \min_{c\in\mathbb{R}} \sum_{i=1}^{n} p_i(u_i - c)^2.$$

Choosing the particular value $c = 0$ gives

$$u^\top H(z)u \leq \sum_{i=1}^{n} p_i u_i^2 \leq \sum_{i=1}^{n} u_i^2 = \|u\|_2^2 ,$$

because each $p_i \in [0, 1]$ and the p_i sum to 1. This proves the upper bound.

The eigenvalue statements follow immediately, because the displayed quadratic bounds hold for every $u \in G$. □

D.4. A precise local cubic remainder

The next proposition addresses the local-ε regime directly and gives the explicit cubic remainder.

Proposition D.6 (Explicit third-order remainder for log-sum-exp)**.** *Let $A(z) = \log \sum_{i=1}^{n} e^{z_i}$. For every $z, h \in \mathbb{R}^n$,*

$$D^3 A(z)[h, h, h] = \sum_{i=1}^{n} p_i(z)\big(h_i - \bar{h}_p\big)^3, \qquad \bar{h}_p := \sum_{i=1}^{n} p_i(z) h_i.$$

Consequently,

$$\big|D^3 A(z)[h, h, h]\big| \leq 2 \|h\|_2^3 .$$

Therefore, for any $u, v \in \mathbb{R}^n$,

$$\left| D_A(v, u) - \frac{1}{2}(v - u)^\top H(u)(v - u) \right| \leq \frac{1}{3} \|v - u\|_2^3 .$$

Proof. Fix $z, h \in \mathbb{R}^n$ and define the one-variable function

$$g(t) := A(z + th), \qquad t \in \mathbb{R}.$$

Let

$$p_t := \mathrm{softmax}(z + th).$$

Then

$$g'(t) = \langle p_t, h \rangle = \sum_{i=1}^{n} (p_t)_i h_i.$$

Define the p_t-mean of the coordinates of h by

$$\mu_t := \sum_{i=1}^{n} (p_t)_i h_i.$$

Thus $g'(t) = \mu_t$.

Differentiating $(p_t)_i$ with respect to t gives

$$\frac{\mathrm{d}}{\mathrm{d}t}(p_t)_i = (p_t)_i \big(h_i - \mu_t\big).$$

Substituting the standard softmax derivative into the derivative of μ_t gives

$$\begin{aligned} g''(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{n} (p_t)_i h_i \\ &= \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}t}(p_t)_i \, h_i \\ &= \sum_{i=1}^{n} (p_t)_i \big(h_i - \mu_t\big) h_i \\ &= \sum_{i=1}^{n} (p_t)_i h_i^2 - \mu_t \sum_{i=1}^{n} (p_t)_i h_i \\ &= \sum_{i=1}^{n} (p_t)_i h_i^2 - \mu_t^2. \end{aligned}$$

This is the variance of the random variable taking value h_i with probability $(p_t)_i$.

Differentiate once more. Because $g''(t) = \sum_i (p_t)_i (h_i - \mu_t)^2$, we obtain

$$\begin{aligned} g'''(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \sum_{i=1}^{n} (p_t)_i (h_i - \mu_t)^2 \\ &= \sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}t}(p_t)_i \, (h_i - \mu_t)^2 + \sum_{i=1}^{n} (p_t)_i \, 2(h_i - \mu_t)(-\mu_t'). \end{aligned}$$

Now $\mu_t' = g''(t)$; its explicit value is unnecessary, because

$$\sum_{i=1}^{n} (p_t)_i (h_i - \mu_t) = 0.$$

Therefore the second sum vanishes, leaving

$$\begin{aligned} g'''(t) &= \sum_{i=1}^{n} (p_t)_i \big(h_i - \mu_t\big)(h_i - \mu_t)^2 \\ &= \sum_{i=1}^{n} (p_t)_i \big(h_i - \mu_t\big)^3. \end{aligned}$$

At $t = 0$ this becomes

$$D^3 A(z)[h,h,h] = g'''(0) = \sum_{i=1}^{n} p_i(z)\big(h_i - \bar{h}_p\big)^3,$$

which proves the formula.

We now bound its magnitude. By the triangle inequality,

$$\left|D^3 A(z)[h,h,h]\right| \le \sum_{i=1}^{n} p_i(z)\left|h_i - \bar{h}_p\right|^3.$$

For every i,

$$|h_i - \bar{h}_p| \le |h_i| + |\bar{h}_p|.$$

Also,

$$|\bar{h}_p| = \left|\sum_{j=1}^{n} p_j(z) h_j\right| \le \sum_{j=1}^{n} p_j(z)|h_j| \le \|h\|_\infty \le \|h\|_2 .$$

Similarly $|h_i| \le \|h\|_\infty \le \|h\|_2$. Hence

$$|h_i - \bar{h}_p| \le 2\,\|h\|_2 \qquad \text{for every } i.$$

Using this inside one factor of the cube gives

$$|h_i - \bar{h}_p|^3 \le 2\,\|h\|_2\ (h_i - \bar{h}_p)^2.$$

Therefore

$$\begin{aligned}
\left|D^3 A(z)[h,h,h]\right| &\le 2\,\|h\|_2 \sum_{i=1}^{n} p_i(z)(h_i - \bar{h}_p)^2 \\
&= 2\,\|h\|_2\ g''(0) \\
&\le 2\,\|h\|_2 \sum_{i=1}^{n} p_i(z) h_i^2 \\
&\le 2\,\|h\|_2\ \|h\|_2^2 \\
&= 2\,\|h\|_2^3 .
\end{aligned}$$

This proves the third-derivative norm bound.

Finally, let $d := v - u$. Taylor's theorem with integral remainder or the mean-value form along the line segment $u + td$ gives

$$A(v) = A(u) + \langle \nabla A(u), d\rangle + \frac{1}{2} d^\top H(u) d + \frac{1}{6} D^3 A(u + \theta d)[d,d,d]$$

for some $\theta \in (0,1)$. Subtracting $A(u) + \langle \nabla A(u), d\rangle$ from both sides yields

$$D_A(v,u) = \frac{1}{2} d^\top H(u) d + \frac{1}{6} D^3 A(u + \theta d)[d,d,d].$$

Taking absolute values and applying the bound already proved gives

$$\left| D_A(v,u) - \frac{1}{2} d^\top H(u) d \right| \le \frac{1}{6} \cdot 2\,\|d\|_2^3 = \frac{1}{3}\,\|d\|_2^3 .$$

Since $d = v - u$, this is exactly the desired remainder estimate. □

Remark D.7. Proposition D.6 gives the local small-ε expansion corresponding to the bounded-logit reduction. If $v = u + \varepsilon\Delta$, then

$$D_A(v,u) = \frac{\varepsilon^2}{2} \Delta^\top H(u) \Delta + O(\varepsilon^3\,\|\Delta\|_2^3),$$

and the explicit constant in front of the cubic term is at most $1/3$. The main theorem below, however, does not rely on $\varepsilon \to 0$; it uses the exact bounded-logit sandwich instead.

E. Bounded-logit reduction from cross-entropy to low-rank approximation

This section proves the main objective theorem.

Theorem E.1 (Bounded-logit cross-entropy induces a weighted matrix-approximation objective)**.** *Let $B > 0$. Assume that the teacher logits satisfy*

$$\eta_*(x) \in K_B \qquad \text{for } \mu\text{-almost every } x,$$

and let $M \in \mathbb{R}^{n\times m}$ satisfy $\mathbf{1}^\top M = 0$ and

$$M\phi(x) \in K_B \qquad \text{for } \mu\text{-almost every } x.$$

Then

$$\begin{aligned}
&\mathcal{R}(M) - \mathcal{R}_* \\
&\geq \tfrac{m_B}{2}\left(\|r\|^2_{L_2(\mu;G)} + \left\|(C-M)\Sigma^{1/2}\right\|^2_{\mathrm{F}}\right), \\
&\mathcal{R}(M) - \mathcal{R}_* \\
&\leq \tfrac{M_B}{2}\left(\|r\|^2_{L_2(\mu;G)} + \left\|(C-M)\Sigma^{1/2}\right\|^2_{\mathrm{F}}\right).
\end{aligned}$$

Equivalently, if $N := M\Sigma^{1/2}$ and $T := C\Sigma^{1/2}$, then

$$\begin{aligned}
&\mathcal{R}(M) - \mathcal{R}_* \\
&\geq \tfrac{m_B}{2}\left(\|r\|^2_{L_2(\mu;G)} + \|T-N\|^2_{\mathrm{F}}\right), \\
&\mathcal{R}(M) - \mathcal{R}_* \\
&\leq \tfrac{M_B}{2}\left(\|r\|^2_{L_2(\mu;G)} + \|T-N\|^2_{\mathrm{F}}\right).
\end{aligned}$$

Proof. By definition,

$$\mathcal{R}(M) = \mathbb{E}_{X,Y\sim p_*}\big[-\log \mathrm{softmax}(M\phi(X))_Y\big].$$

Condition first on X. For each fixed context x, the conditional law of Y is $p_*(\cdot \mid x)$. Therefore

$$\mathcal{R}(M) = \mathbb{E}_X\Big[\sum_{i=1}^n p_*(i \mid X)\big(-\log \mathrm{softmax}(M\phi(X))_i\big)\Big].$$

Similarly,

$$\mathcal{R}_* = \mathbb{E}_X\Big[\sum_{i=1}^n p_*(i \mid X)\big(-\log p_*(i \mid X)\big)\Big].$$

Subtract the two formulas. We obtain

$$\begin{aligned}
\mathcal{R}(M) - \mathcal{R}_* &= \mathbb{E}_X\Big[\sum_{i=1}^n p_*(i \mid X)\log\frac{p_*(i \mid X)}{\mathrm{softmax}(M\phi(X))_i}\Big] \\
&= \mathbb{E}_X\Big[\mathrm{KL}\left(p_*(\cdot \mid X) \,\|\, \mathrm{softmax}(M\phi(X))\right)\Big].
\end{aligned}$$

Because $p_*(\cdot \mid X) = \mathrm{softmax}(\eta_*(X))$, Lemma D.3 yields

$$\mathrm{KL}\left(p_*(\cdot \mid X) \,\|\, \mathrm{softmax}(M\phi(X))\right) = D_A\big(M\phi(X), \eta_*(X)\big).$$

Hence

$$\mathcal{R}(M) - \mathcal{R}_* = \mathbb{E}_X\Big[D_A\big(M\phi(X), \eta_*(X)\big)\Big].$$

On the bounded logit box, that Bregman divergence is uniformly comparable to a quadratic form.

Fix a context x. Both $\eta_*(x)$ and $M\phi(x)$ lie in K_B by assumption, and K_B is convex, so the entire segment between them also lies in K_B. Therefore Lemma D.4 and the definitions of m_B, M_B imply

$$\begin{aligned}\tfrac{m_B}{2}\left\|M\phi(x)-\eta_*(x)\right\|_2^2 &\le D_A\big(M\phi(x),\eta_*(x)\big)\\ &\le \tfrac{M_B}{2}\left\|M\phi(x)-\eta_*(x)\right\|_2^2.\end{aligned}$$

Take expectation over X to obtain

$$\begin{aligned}\tfrac{m_B}{2}\left\|M\phi-\eta_*\right\|_{L_2(\mu;G)}^2 &\le \mathcal{R}(M)-\mathcal{R}_*\\ &\le \tfrac{M_B}{2}\left\|M\phi-\eta_*\right\|_{L_2(\mu;G)}^2.\end{aligned}$$

It remains to identify the $L_2(\mu;G)$ norm of the logit error relative to the projected teacher.

Recall that

$$\eta_* = C\phi + r, \qquad r \perp \mathcal{F}.$$

Also, because both $C\phi$ and $M\phi$ belong to $\mathcal{F}$, their difference

$$(C-M)\phi = C\phi - M\phi$$

also belongs to $\mathcal{F}$. Therefore r is orthogonal to $(C-M)\phi$ in $L_2(\mu;G)$. Now

$$\begin{aligned}M\phi-\eta_* &= M\phi-(C\phi+r)\\ &= (M-C)\phi - r.\end{aligned}$$

Since $r \perp (M-C)\phi$, the Pythagorean theorem in the Hilbert space $L_2(\mu;G)$ gives

$$\left\|M\phi-\eta_*\right\|_{L_2(\mu;G)}^2 = \left\|r\right\|_{L_2(\mu;G)}^2 + \left\|(C-M)\phi\right\|_{L_2(\mu;G)}^2.$$

The remaining term is exactly a Frobenius norm.

Set $A_M := C - M$. Then

$$\left\|(C-M)\phi\right\|_{L_2(\mu;G)}^2 = \mathbb{E}\left\|A_M\phi(X)\right\|_2^2.$$

Expand the squared Euclidean norm:

$$\begin{aligned}\mathbb{E}\left\|A_M\phi(X)\right\|_2^2 &= \mathbb{E}\big[(A_M\phi(X))^\top(A_M\phi(X))\big]\\ &= \mathbb{E}\big[\operatorname{tr}\big((A_M\phi(X))^\top(A_M\phi(X))\big)\big]\\ &= \mathbb{E}\big[\operatorname{tr}\big(A_M\phi(X)\phi(X)^\top A_M^\top\big)\big]\\ &= \operatorname{tr}\Big(A_M\,\mathbb{E}[\phi(X)\phi(X)^\top]A_M^\top\Big)\\ &= \operatorname{tr}\big(A_M\Sigma A_M^\top\big)\\ &= \left\|A_M\Sigma^{1/2}\right\|_{\mathrm{F}}^2\\ &= \left\|(C-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.\end{aligned}$$

Therefore

$$\left\|M\phi-\eta_*\right\|_{L_2(\mu;G)}^2 = \left\|r\right\|_{L_2(\mu;G)}^2 + \left\|(C-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

Substituting this identity into the quadratic comparison above yields

$$\begin{aligned}&\mathcal{R}(M)-\mathcal{R}_*\\ &\ge \tfrac{m_B}{2}\left(\left\|r\right\|_{L_2(\mu;G)}^2 + \left\|(C-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2\right),\\ &\mathcal{R}(M)-\mathcal{R}_*\\ &\le \tfrac{M_B}{2}\left(\left\|r\right\|_{L_2(\mu;G)}^2 + \left\|(C-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2\right).\end{aligned}$$

This is the first displayed formula of the theorem.

The second displayed formula follows by the change of variables. Since $T = C\Sigma^{1/2}$ and $N = M\Sigma^{1/2}$,

$$(C - M)\Sigma^{1/2} = T - N.$$

Substituting this into the previous display yields the equivalent form. □

Remark E.2 (Interpretation). Theorem E.1 provides the reduction from next-token cross-entropy to pre-linearization. It says that, on a bounded logit region, population cross-entropy is not merely *approximately* but *uniformly comparable* to a matrix-approximation objective built from the projected teacher logit map. The matrix $T = C\Sigma^{1/2}$ is the key object. It combines two effects:

- the dictionary-to-logit coefficient matrix C, which measures how useful each interaction direction is for prediction, and
- the feature second-moment factor $\Sigma^{1/2}$, which weights interaction directions by how much they occur under the data distribution.

Therefore large singular directions of T are exactly the interaction directions that are simultaneously *common* and *predictively important*.

F. Exact solution of the induced width-k problem

The surrogate problem induced by Theorem E.1 has a closed-form SVD solution.

Theorem F.1 (Exact width-k solution by SVD)**.** *Assume $\Sigma \succ 0$. Let*

$$T = C\Sigma^{1/2} \in \mathbb{R}^{n\times m}$$

and let its singular value decomposition be

$$T = U\ \mathrm{diag}(s_1, \dots, s_q, 0, \dots, 0)\, V^\top, \qquad s_1 \geq s_2 \geq \cdots \geq s_q > 0,$$

where $q = \mathrm{rank}(T)$. For $0 \leq k \leq q$, define the surrogate objective

$$\mathcal{E}(M) := \left\|(C - M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2 = \left\|T - M\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

Then

$$\min_{\mathrm{rank}(M)\leq k} \mathcal{E}(M) = \sum_{i=k+1}^{q} s_i^2.$$

An optimizer is given by

$$M_k := U_k\ \mathrm{diag}(s_1, \dots, s_k)\, V_k^\top \Sigma^{-1/2},$$

where U_k and V_k contain the first k singular vectors. Equivalently, the optimal width-k representation is

$$h_k(x) := V_k^\top \Sigma^{-1/2}\phi(x) \in \mathbb{R}^k,$$

with readout matrix

$$U_k\ \mathrm{diag}(s_1, \dots, s_k).$$

Thus the optimal width-k student keeps a top-k singular subspace of the whitened interaction-to-logit operator T. If $s_k > s_{k+1}$, this subspace is unique; if there is a tie at the cutoff, any top-k subspace inside the tied singular space is optimal.

Proof. Because $\Sigma \succ 0$, the matrix $\Sigma^{1/2}$ is invertible. Therefore the map

$$M \longmapsto N := M\Sigma^{1/2}$$

is a bijection from $\mathbb{R}^{n\times m}$ to itself, and it preserves rank because multiplication by an invertible matrix on the right does not change the row rank or the full matrix rank. Hence

$$\min_{\mathrm{rank}(M)\leq k} \mathcal{E}(M) = \min_{\mathrm{rank}(N)\leq k} \|T - N\|_{\mathrm{F}}^2.$$

So it is enough to solve the latter problem.

Let $N \in \mathbb{R}^{n\times m}$ with $\operatorname{rank}(N) \le k$, and let $S \subseteq \mathbb{R}^m$ be the row space of N. Then $\dim S \le k$. Let P_S denote the orthogonal projector onto S. For this fixed row space, the best approximation is the orthogonal projection of T onto S.

Because every row of N lies in S, right multiplication by P_S leaves N unchanged:

$$N = NP_S.$$

Therefore

$$T - N = T - NP_S = T(I - P_S) + TP_S - N.$$

We claim that the two matrices on the right-hand side are Frobenius-orthogonal. Indeed,

$$\begin{aligned}\langle T(I-P_S), TP_S - N\rangle_{\mathrm{F}} &= \operatorname{tr}\big((T(I-P_S))^\top (TP_S - N)\big)\\ &= \operatorname{tr}\big((I-P_S)T^\top (TP_S - N)\big).\end{aligned}$$

Now use cyclicity of the trace to move the left factor to the end:

$$\operatorname{tr}\big((I-P_S)T^\top (TP_S - N)\big) = \operatorname{tr}\big(T^\top (TP_S - N)(I-P_S)\big).$$

But $TP_S(I - P_S) = 0$ because $P_S(I - P_S) = 0$, and also

$$N(I - P_S) = NP_S(I - P_S) = 0.$$

Thus the whole trace is zero. Therefore

$$\|T - N\|_{\mathrm{F}}^2 = \|T(I - P_S)\|_{\mathrm{F}}^2 + \|TP_S - N\|_{\mathrm{F}}^2 \ge \|T(I - P_S)\|_{\mathrm{F}}^2 .$$

Equality holds if and only if $TP_S - N = 0$, that is, if and only if $N = TP_S$.

So, once the row subspace S is fixed, the best rank-at-most-k approximation supported on that row subspace is $N = TP_S$, and the resulting error is $\|T(I - P_S)\|_{\mathrm{F}}^2$.

This reduces the problem to an optimization over k-dimensional subspaces.

The previous step shows that

$$\min_{\operatorname{rank}(N)\le k} \|T - N\|_{\mathrm{F}}^2 = \min_{\dim S\le k} \|T(I - P_S)\|_{\mathrm{F}}^2 .$$

Since

$$\|T\|_{\mathrm{F}}^2 = \|TP_S\|_{\mathrm{F}}^2 + \|T(I - P_S)\|_{\mathrm{F}}^2$$

by the same orthogonality argument, minimizing $\|T(I - P_S)\|_{\mathrm{F}}^2$ is equivalent to maximizing $\|TP_S\|_{\mathrm{F}}^2$. Therefore

$$\min_{\operatorname{rank}(N)\le k} \|T - N\|_{\mathrm{F}}^2 = \|T\|_{\mathrm{F}}^2 - \max_{\dim S\le k} \|TP_S\|_{\mathrm{F}}^2 .$$

Now compute

$$\|TP_S\|_{\mathrm{F}}^2 = \operatorname{tr}\big((TP_S)^\top (TP_S)\big) = \operatorname{tr}\big(P_S T^\top TP_S\big) = \operatorname{tr}\big(P_S T^\top T\big),$$

because $P_S^2 = P_S$ and the trace is cyclic. So the problem becomes

$$\max_{\dim S\le k} \operatorname{tr}(P_S T^\top T).$$

Diagonalizing $T^\top T$ then solves the projection problem exactly.

Because $T = U \operatorname{diag}(s_1, \dots, s_q, 0, \dots, 0)V^\top$, we have

$$T^\top T = V \operatorname{diag}(s_1^2, \dots, s_q^2, 0, \dots, 0)V^\top.$$

Let P be any orthogonal projector of rank at most k. Define

$$Q := V^\top PV.$$

Then Q is an orthogonal projector of rank at most k; in particular, Q is symmetric, $Q^2 = Q$, $0 \le Q_{ii} \le 1$, and $\sum_i Q_{ii} = \mathrm{tr}(Q) = \mathrm{rank}(Q) \le k$. Now

$$\begin{aligned}\mathrm{tr}(PT^\top T) &= \mathrm{tr}\left(V^\top PV\ \mathrm{diag}(s_1^2, \dots, s_m^2)\right)\\ &= \mathrm{tr}\left(Q\ \mathrm{diag}(s_1^2, \dots, s_m^2)\right)\\ &= \sum_{i=1}^m s_i^2 Q_{ii}.\end{aligned}$$

The vector of diagonal entries $(Q_{11}, \dots, Q_{mm})$ lies in the polytope

$$\Big\{q \in [0,1]^m : \sum_i q_i \le k\Big\}.$$

The objective $\sum_i s_i^2 q_i$ is linear in q, and the coefficients are ordered as

$$s_1^2 \ge s_2^2 \ge \cdots \ge s_m^2 \ge 0.$$

A linear function on a compact polytope attains its maximum at an extreme point. The extreme points of the displayed polytope are vectors with entries in $\{0,1\}$ and at most k ones. Since the coefficients are sorted in descending order, the maximum is attained by taking $q_i = 1$ for $i \le k$ and $q_i = 0$ for $i > k$. Therefore

$$\mathrm{tr}(PT^\top T) \le \sum_{i=1}^k s_i^2.$$

Equality is achieved by the projector

$$P_* := V_k V_k^\top.$$

Hence

$$\max_{\dim S \le k} \|TP_S\|_{\mathrm{F}}^2 = \sum_{i=1}^k s_i^2.$$

It follows that

$$\min_{\mathrm{rank}(N) \le k} \|T - N\|_{\mathrm{F}}^2 = \|T\|_{\mathrm{F}}^2 - \sum_{i=1}^k s_i^2 = \sum_{i=k+1}^q s_i^2.$$

This proves the value of the minimum.

It only remains to identify an optimizer and rewrite it as a width-k representation.

The row-space argument above shows that the best approximation associated with the optimal projector $P_* = V_k V_k^\top$ is

$$N_k = TP_* = TV_k V_k^\top.$$

Using the SVD of T gives

$$N_k = U_k\, \mathrm{diag}(s_1, \dots, s_k) V_k^\top.$$

Since $N_k = M_k \Sigma^{1/2}$, we solve for M_k:

$$M_k = N_k \Sigma^{-1/2} = U_k\, \mathrm{diag}(s_1, \dots, s_k) V_k^\top \Sigma^{-1/2}.$$

This is exactly the formula stated in the theorem. Since $\mathbf{1}^\top T = 0$, every nonzero left singular vector of T lies in G, so $\mathbf{1}^\top M_k = 0$ and the optimizer respects the zero-sum gauge.

Now define

$$h_k(x) := V_k^\top \Sigma^{-1/2} \phi(x) = V_k^\top \widetilde{\phi}(x) \in \mathbb{R}^k.$$

Then

$$\begin{aligned}M_k \phi(x) &= U_k\, \mathrm{diag}(s_1, \dots, s_k) V_k^\top \Sigma^{-1/2} \phi(x)\\ &= U_k\, \mathrm{diag}(s_1, \dots, s_k) h_k(x).\end{aligned}$$

So the optimizer indeed factors through the width-k representation h_k. □

Remark F.2 (Interpretation). Theorem F.1 says that the cross-entropy-induced surrogate problem need not be solved by selecting named dictionary coordinates. It is solved by keeping the top singular directions of the matrix $T = C\Sigma^{1/2}$. In general, those singular directions are linear combinations of the original dictionary features. Only under additional structure do they coincide with individual named interactions. We study that case next.

G. Orthogonal interactions, commonness, and named-feature selection

G.1. Exact coordinate selection in the orthogonal case

We specialize to the case where the interaction dictionary is orthogonal in $L_2(\mu)$ and each interaction affects the logits in an orthogonal output direction.

Assumption G.1 (Orthogonal-coordinate regime)**.** Assume that the dictionary is $\phi = (\psi_1, \dots, \psi_m)$ with

$$\mathbb{E}[\psi_r \psi_s] = 0 \quad (r \neq s), \qquad \mathbb{E}[\psi_r^2] = \nu_r > 0.$$

Thus $\Sigma = \mathrm{diag}(\nu_1, \dots, \nu_m)$. Write the columns of C as

$$C = [c_1 \; c_2 \; \cdots \; c_m], \qquad c_r \in G.$$

Assume also that these output directions are pairwise orthogonal in $\mathbb{R}^n$:

$$\langle c_r, c_s \rangle = 0 \quad (r \neq s).$$

Define the score of coordinate r by

$$d_r := \nu_r \, \|c_r\|_2^2 \, .$$

Without loss of generality, order the coordinates so that

$$d_1 \geq d_2 \geq \cdots \geq d_m.$$

Theorem G.2 (Exact named-feature selection under orthogonality)**.** *Under Assumption G.1, the optimal width-k surrogate representation is exactly the span of the first k named interaction coordinates. More precisely,*

$$\min_{\mathrm{rank}(M) \leq k} \left\| (C - M)\Sigma^{1/2} \right\|_{\mathrm{F}}^2 = \sum_{r=k+1}^{m} d_r,$$

and an optimizer is obtained by keeping precisely the first k columns of C and setting the remaining columns to zero. Equivalently, the optimal width-k representation can be taken to be

$$h_k(x) = (\psi_1(x), \dots, \psi_k(x))$$

up to coordinate-wise rescaling.

Proof. Because $\Sigma = \mathrm{diag}(\nu_1, \dots, \nu_m)$, the whitened operator $T = C\Sigma^{1/2}$ has columns

$$\sqrt{\nu_1} c_1, \; \sqrt{\nu_2} c_2, \; \dots, \; \sqrt{\nu_m} c_m.$$

By the orthogonality assumption on the c_r, these columns are pairwise orthogonal. Hence

$$T^\top T = \mathrm{diag}(\nu_1 \|c_1\|_2^2, \dots, \nu_m \|c_m\|_2^2) = \mathrm{diag}(d_1, \dots, d_m).$$

So the right singular vectors of T are exactly the standard basis vectors $e_1, \dots, e_m$, and the singular values are

$$s_r = \sqrt{d_r}.$$

Because the d_r are arranged in descending order, the top k right singular directions are exactly $e_1, \dots, e_k$. Therefore Theorem F.1 applies with

$$\sum_{i=k+1}^{q} s_i^2 = \sum_{r=k+1}^{m} d_r.$$

This proves the formula for the minimum surrogate error.

To identify the optimizer explicitly, note that keeping the first k right singular directions means keeping the first k coordinates of the whitened dictionary and discarding the rest. In the original unwhitened coordinates, that corresponds to using the representation

$$(\psi_1/\sqrt{\nu_1}, \dots, \psi_k/\sqrt{\nu_k})$$

up to an invertible diagonal rescaling. Since linear readability is unchanged by invertible rescaling of the representation coordinates, this is equivalent to using

$$h_k(x) = (\psi_1(x), \dots, \psi_k(x)).$$

The corresponding coefficient matrix keeps the first k columns of C and sets the remaining columns to zero. This is the best rank-k approximation because the columns are orthogonal and ordered by squared norm. That completes the proof. □

Remark G.3. Theorem G.2 formalizes, in the orthogonal-coordinate case, the principle that common and useful features are pre-linearized first. The commonness factor is ν_r, while the usefulness factor is $\|c_r\|_2^2$.

G.2. Indicator features and prevalence

Corollary G.4 (Indicator interactions)**.** *Let $\psi_r = \mathbf{1}_{E_r} - p_r$, where E_r is an event and $p_r = \mu(E_r)$. Then*

$$\nu_r = \mathbb{E}[\psi_r^2] = p_r(1 - p_r).$$

In particular, if $p_r \le \frac{1}{2}$, then

$$\frac{p_r}{2} \le \nu_r \le p_r.$$

Thus, for rare events, the score $d_r = \nu_r \|c_r\|_2^2$ is within a factor of 2 of

$$p_r \|c_r\|_2^2 .$$

The ranking is therefore prevalence times predictive usefulness up to this constant factor.

Proof. By definition,

$$\psi_r = \mathbf{1}_{E_r} - p_r.$$

Therefore

$$\begin{aligned}
\nu_r &= \mathbb{E}[(\mathbf{1}_{E_r} - p_r)^2] \\
&= \mathbb{E}[\mathbf{1}_{E_r}^2] - 2p_r\mathbb{E}[\mathbf{1}_{E_r}] + p_r^2 \\
&= p_r - 2p_r^2 + p_r^2 \\
&= p_r(1 - p_r).
\end{aligned}$$

If $p_r \le \frac{1}{2}$, then $1 - p_r \in [\frac{1}{2}, 1]$. Multiplying by p_r gives

$$\frac{p_r}{2} \le p_r(1 - p_r) \le p_r.$$

This proves the claim. □

G.3. A concrete conjunction corollary

Corollary G.5 (Pairwise conjunctions of binary primitive features)**.** *Let $x_i, x_j \in \{0, 1\}$ be independent Bernoulli variables with means p_i, p_j. Define the centered interaction*

$$\psi_{ij} := (x_i - p_i)(x_j - p_j).$$

Then

$$\mathbb{E}[\psi_{ij}] = 0, \qquad \mathbb{E}[\psi_{ij}^2] = p_i(1 - p_i)p_j(1 - p_j).$$

Moreover,

$$x_i x_j = \psi_{ij} + p_j x_i + p_i x_j - p_i p_j.$$

Therefore, if ψ_{ij}, x_i, x_j, and 1 are linearly readable from a representation, then $x_i \wedge x_j = x_i x_j$ is also linearly readable from that representation. In the sparse regime $p_i, p_j \ll 1$, the variance factor is approximately

$$p_i p_j = \Pr[x_i = 1, x_j = 1],$$

which is exactly the prevalence of the conjunction event.

Proof. Because x_i and x_j are independent, we have

$$\mathbb{E}[x_i - p_i] = 0, \qquad \mathbb{E}[x_j - p_j] = 0,$$

and hence

$$\mathbb{E}[\psi_{ij}] = \mathbb{E}[x_i - p_i]\,\mathbb{E}[x_j - p_j] = 0.$$

Also,

$$\begin{aligned}\mathbb{E}[\psi_{ij}^2] &= \mathbb{E}[(x_i - p_i)^2 (x_j - p_j)^2] \\ &= \mathbb{E}[(x_i - p_i)^2]\mathbb{E}[(x_j - p_j)^2]\end{aligned}$$

by independence. For a Bernoulli random variable $x \sim \mathrm{Bernoulli}(p)$,

$$\mathbb{E}[(x - p)^2] = p(1 - p).$$

Therefore

$$\mathbb{E}[\psi_{ij}^2] = p_i(1 - p_i)p_j(1 - p_j).$$

This proves the first part.

For the second part, expand the definition of ψ_{ij}:

$$\begin{aligned}\psi_{ij} &= (x_i - p_i)(x_j - p_j) \\ &= x_i x_j - p_j x_i - p_i x_j + p_i p_j.\end{aligned}$$

Rearranging gives

$$x_i x_j = \psi_{ij} + p_j x_i + p_i x_j - p_i p_j.$$

Thus if x_i, x_j, and ψ_{ij} are linearly readable, then $x_i x_j$ is also linearly readable. Since $x_i \wedge x_j = x_i x_j$ for binary variables, the conjunction becomes linearly readable as well.

Finally, in the sparse regime $p_i, p_j \ll 1$, each factor $(1 - p_i)$ and $(1 - p_j)$ is close to 1. Therefore

$$p_i(1 - p_i)p_j(1 - p_j) \approx p_i p_j,$$

and because of independence,

$$p_i p_j = \Pr[x_i = 1]\Pr[x_j = 1] = \Pr[x_i = 1, x_j = 1].$$

That is the prevalence of the conjunction event. □

H. General case: subspaces, not named features

The orthogonal-coordinate theorem gives exact coordinate selection. Without orthogonality, interaction features may push probability mass toward overlapping token groups, so the selected object is generally a subspace rather than a coordinate list.

We first restate and prove Theorem 3.3, which asserts that the selected whitened subspace is invariant under any invertible reparameterization of the dictionary.

Theorem H.1 (Basis invariance of the selected whitened subspace, restated)**.** *Let $\phi' = A\phi$ for some invertible matrix $A \in \mathbb{R}^{m\times m}$, and define*

$$\Sigma' := \mathbb{E}[\phi'(X)\phi'(X)^\top] = A\Sigma A^\top.$$

Let C' satisfy $C'\phi' = C\phi$ and set $T' := C'\Sigma'^{1/2}$. Then

$$Q := \Sigma'^{-1/2} A \Sigma^{1/2}$$

is orthogonal and

$$T' = TQ^\top.$$

Consequently T and T' have the same singular values. If V_k spans the top-k right singular subspace of T, then QV_k spans the top-k right singular subspace of T', and the readable scalar space

$$\operatorname{span}\{v^\top \Sigma^{-1/2}\phi(\cdot) : v \in \operatorname{range}(V_k)\}$$

is exactly equal to

$$\operatorname{span}\{v'^\top \Sigma'^{-1/2}\phi'(\cdot) : v' \in \operatorname{range}(QV_k)\}.$$

Thus the canonical object is a basis-invariant whitened interaction subspace, not a basis-dependent coordinate list.

Proof. Because $\phi' = A\phi$, the whitened coordinates satisfy

$$\widetilde{\phi}' := \Sigma'^{-1/2}\phi' = \Sigma'^{-1/2} A \Sigma^{1/2}\widetilde{\phi} = Q\widetilde{\phi}.$$

Moreover,

$$QQ^\top = \Sigma'^{-1/2} A\Sigma A^\top \Sigma'^{-1/2} = \Sigma'^{-1/2}\Sigma'\Sigma'^{-1/2} = I,$$

so Q is orthogonal.

Since $C'\phi' = C\phi$ almost surely, we have $(C'A - C)\phi(X) = 0$ almost surely. Hence

$$0 = \mathbb{E}\,\|(C'A - C)\phi(X)\|_2^2 = \operatorname{tr}\big((C'A - C)\Sigma(C'A - C)^\top\big).$$

Because $\Sigma \succ 0$, this implies $C'A = C$. Therefore

$$T'Q = C'\Sigma'^{1/2}\Sigma'^{-1/2} A\Sigma^{1/2} = C'A\Sigma^{1/2} = C\Sigma^{1/2} = T,$$

which is equivalent to $T' = TQ^\top$.

If $T = U\operatorname{Diag}(s)V^\top$, then

$$T' = U\operatorname{Diag}(s)(QV)^\top,$$

so T and T' have the same singular values and the right singular vectors rotate by Q. Finally, if $v' = Qv$, then

$$v'^\top \widetilde{\phi}' = v^\top Q^\top Q\widetilde{\phi} = v^\top \widetilde{\phi}.$$

Hence the scalar readable space generated by the selected singular directions is unchanged. □

Having established invariance, we next quantify when the true cross-entropy optimum remains close to the surrogate-selected whitened subspace.

Theorem H.2 (Approximate recovery of the surrogate-optimal interaction subspace from actual cross-entropy)**.** *Assume $\Sigma \succ 0$ and $r = 0$, so that the teacher logit map lies exactly in the dictionary family:*

$$\eta_* = C\phi.$$

Let

$$T = C\Sigma^{1/2} = U\operatorname{diag}(s_1, \ldots, s_q, 0, \ldots, 0)V^\top$$

with singular values ordered as

$$s_1 \geq s_2 \geq \cdots \geq s_q > 0.$$

Fix $k < q$ and assume there is a nonzero spectral gap

$$g := s_k^2 - s_{k+1}^2 > 0.$$

Let

$$P_* := V_k V_k^\top$$

be the orthogonal projector onto the top-k right singular subspace of T, and define the optimal surrogate error

$$E_* := \sum_{i=k+1}^{q} s_i^2.$$

Let M_{CE} be any rank-at-most-k minimizer of the true cross-entropy risk among admissible students that remain in K_B, and let P_{CE} be the orthogonal projector onto the row space of $M_{\mathrm{CE}}\Sigma^{1/2}$. Assume also that the surrogate-optimal width-k student M_k from Theorem F.1 is admissible in the same bounded-logit class. Then

$$\|P_{\mathrm{CE}} - P_*\|_{\mathrm{F}}^2 \leq \frac{2(\kappa_B - 1)E_*}{g}.$$

In particular, when the bounded-logit condition number κ_B is close to 1 and the gap g is large relative to the tail energy E_*, the true cross-entropy optimum has small Frobenius projector distance from the surrogate-optimal singular subspace.

Proof. Because $r = 0$, Theorem E.1 simplifies to

$$\tfrac{m_B}{2}\mathcal{E}(M) \leq \mathcal{R}(M) - \mathcal{R}_* \leq \tfrac{M_B}{2}\mathcal{E}(M),$$
$$\mathcal{E}(M) := \left\|(C - M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

Apply the lower bound to M_{CE} and the upper bound to the surrogate-optimal student M_k. Since M_{CE} minimizes the true risk over the admissible rank-at-most-k class, we have

$$\mathcal{R}(M_{\mathrm{CE}}) \leq \mathcal{R}(M_k).$$

Therefore

$$\frac{m_B}{2}\mathcal{E}(M_{\mathrm{CE}}) \leq \mathcal{R}(M_{\mathrm{CE}}) - \mathcal{R}_* \leq \mathcal{R}(M_k) - \mathcal{R}_* \leq \frac{M_B}{2}\mathcal{E}(M_k).$$

Multiply by $2/m_B$ to obtain

$$\mathcal{E}(M_{\mathrm{CE}}) \leq \frac{M_B}{m_B}\mathcal{E}(M_k) = \kappa_B E_*.$$

Here we used Theorem F.1, which says that $\mathcal{E}(M_k) = E_*$. Thus

$$\mathcal{E}(M_{\mathrm{CE}}) - E_* \leq (\kappa_B - 1)E_*.$$

Let

$$N_{\mathrm{CE}} := M_{\mathrm{CE}}\Sigma^{1/2}.$$

Fix the row-space projector P_{CE}. The proof of Theorem F.1 shows that the best approximation with this row space is TP_{CE}, with error $\|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2$. Since N_{CE} has this row space and rank at most k, we have

$$\|T - N_{\mathrm{CE}}\|_{\mathrm{F}}^2 \geq \|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2.$$

Since $\mathcal{E}(M_{\mathrm{CE}}) = \|T - N_{\mathrm{CE}}\|_{\mathrm{F}}^2$, this implies

$$\|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2 \leq \mathcal{E}(M_{\mathrm{CE}}) \leq \kappa_B E_*.$$

Thus

$$\|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2 - E_* \leq (\kappa_B - 1)E_*.$$

We now lower-bound the excess surrogate error by the subspace misalignment without assuming that P_{CE} has rank exactly k.

Let

$$D := \mathrm{diag}(s_1^2, \ldots, s_m^2).$$

Because $T^\top T = VDV^\top$, define

$$Q := V^\top P_{\mathrm{CE}} V.$$

Then Q is an orthogonal projector of rank

$$r_{\mathrm{CE}} := \mathrm{rank}(P_{\mathrm{CE}}) \leq k.$$

Its diagonal entries satisfy $0 \leq Q_{ii} \leq 1$, and

$$\sum_{i=1}^{m} Q_{ii} = \mathrm{tr}(Q) = r_{\mathrm{CE}}.$$

Now

$$\begin{aligned}\|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2 &= \mathrm{tr}\left((I - P_{\mathrm{CE}})T^\top T\right) \\ &= \mathrm{tr}\left((I - Q)D\right) \\ &= \sum_{i=1}^{m} s_i^2(1 - Q_{ii}).\end{aligned}$$

The optimal surrogate error is

$$E_* = \sum_{i=k+1}^{m} s_i^2.$$

Define the surrogate excess for this row-space projector by

$$\Delta := \|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2 - E_*.$$

Then

$$\Delta = \sum_{i=1}^{k} s_i^2(1 - Q_{ii}) - \sum_{i=k+1}^{m} s_i^2 Q_{ii}.$$

Define

$$\alpha := \sum_{i=k+1}^{m} Q_{ii}, \qquad \beta := \sum_{i=1}^{k}(1 - Q_{ii}).$$

Since $\sum_i Q_{ii} = r_{\mathrm{CE}}$, we have

$$\beta = k - \sum_{i=1}^{k} Q_{ii} = k - (r_{\mathrm{CE}} - \alpha) = k - r_{\mathrm{CE}} + \alpha.$$

Thus $\beta - \alpha = k - r_{\mathrm{CE}} \geq 0$. Using the ordering of the singular values,

$$\Delta \geq s_k^2\beta - s_{k+1}^2\alpha.$$

Moreover, with $g = s_k^2 - s_{k+1}^2$,

$$s_k^2\beta - s_{k+1}^2\alpha - \frac{g}{2}(\beta + \alpha) = \frac{s_k^2 + s_{k+1}^2}{2}(\beta - \alpha) \geq 0.$$

Therefore

$$\Delta \geq \frac{g}{2}(\beta + \alpha).$$

The projector distance has exactly the same $\beta + \alpha$ form:

$$\begin{aligned}
\|P_{\mathrm{CE}} - P_*\|_{\mathrm{F}}^2 &= \operatorname{tr}(P_{\mathrm{CE}}) + \operatorname{tr}(P_*) - 2\operatorname{tr}(P_{\mathrm{CE}}P_*) \\
&= r_{\mathrm{CE}} + k - 2\sum_{i=1}^{k} Q_{ii} \\
&= r_{\mathrm{CE}} + k - 2(r_{\mathrm{CE}} - \alpha) \\
&= k - r_{\mathrm{CE}} + 2\alpha \\
&= \beta + \alpha.
\end{aligned}$$

Consequently,

$$\Delta \geq \frac{g}{2}\|P_{\mathrm{CE}} - P_*\|_{\mathrm{F}}^2.$$

Combining this lower bound with $\Delta \leq (\kappa_B - 1)E_*$ gives

$$\|P_{\mathrm{CE}} - P_*\|_{\mathrm{F}}^2 \leq \frac{2(\kappa_B - 1)E_*}{g}.$$

That is exactly the claimed inequality. □

Remark H.3 (Interpretation). Theorem H.2 formalizes the following point. In the fully general case, the optimal width-k model does not classify individual coordinates as pre-linearized; it selects a *k-dimensional interaction subspace*. Named-feature selection is a special case that arises when this singular subspace aligns with the coordinate axes of the chosen dictionary.

Corollary H.4 (Approximate named-feature recovery in the orthogonal-coordinate case)**.** *Under Assumption G.1, suppose in addition that $r = 0$, that the top-k score gap is positive,*

$$g = d_k - d_{k+1} > 0,$$

and that the surrogate optimizer is admissible in the bounded-logit class. Let P_{CE} be the projector from Theorem H.2. Then the total leakage of the true cross-entropy optimum outside the top-k feature span satisfies

$$\sum_{r=k+1}^{m} \|P_{\mathrm{CE}}e_r\|_2^2 \leq \frac{(\kappa_B - 1)\sum_{r=k+1}^{m} d_r}{d_k - d_{k+1}}.$$

Proof. In the orthogonal-coordinate case, the optimal singular subspace is the coordinate span of $e_1, \ldots, e_k$, so $P_* = \operatorname{diag}(I_k, 0)$. Write

$$\alpha := \sum_{r=k+1}^{m} \|P_{\mathrm{CE}}e_r\|_2^2 = \sum_{r=k+1}^{m} e_r^\top P_{\mathrm{CE}} e_r.$$

This is the same leakage quantity denoted α in the proof of Theorem H.2, specialized to the coordinate basis. That proof gives the direct excess bound

$$\|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2 - E_* \geq (d_k - d_{k+1})\alpha.$$

The bounded-logit comparison gives

$$\|T(I - P_{\mathrm{CE}})\|_{\mathrm{F}}^2 - E_* \leq (\kappa_B - 1)E_*.$$

In the orthogonal-coordinate case $E_* = \sum_{r=k+1}^{m} d_r$, so

$$\alpha \leq \frac{(\kappa_B - 1)\sum_{r=k+1}^{m} d_r}{d_k - d_{k+1}}.$$

This is exactly the claimed inequality. □

I. Stylized dynamics of singular-mode emergence

The preceding theorems explain the objective that cross-entropy induces on a bottlenecked interaction dictionary. We next analyze a stylized gradient-flow model for the induced low-rank problem.

The theorem below concerns gradient flow for the factorized surrogate

$$L(A, B) := \frac{1}{2} \|T - AB\|_{\mathrm{F}}^2 .$$

The result applies only to gradient flow on the factorized surrogate $L(A, B)$; it is not a statement about end-to-end SGD in arbitrary deep networks. It characterizes the rank-factorized matrix problem induced by the objective reduction of Theorem E.1.

Theorem I.1 (Mode-wise gradient flow for the factorized surrogate)**.** *Let*

$$T = \sum_{i=1}^{r} s_i u_i v_i^\top, \qquad s_1 \geq s_2 \geq \cdots \geq s_r > 0,$$

be a singular value decomposition of a rank-r matrix $T \in \mathbb{R}^{n\times m}$. Consider the factorized objective

$$L(A, B) := \frac{1}{2} \|T - AB\|_{\mathrm{F}}^2 , \qquad A \in \mathbb{R}^{n\times r}, \quad B \in \mathbb{R}^{r\times m},$$

with gradient-flow dynamics

$$\dot{A} = -(AB - T)B^\top, \qquad \dot{B} = -A^\top(AB - T).$$

Assume the initialization is singular-vector aligned:

$$A(0) = \sum_{i=1}^{r} a_i(0) u_i e_i^\top, \qquad B(0) = \sum_{i=1}^{r} b_i(0) e_i v_i^\top,$$

with $a_i(0), b_i(0) \in \mathbb{R}$. Then for all time for which the solution exists,

$$A(t) = \sum_{i=1}^{r} a_i(t) u_i e_i^\top, \qquad B(t) = \sum_{i=1}^{r} b_i(t) e_i v_i^\top,$$

and the scalar coefficients satisfy the decoupled ODEs

$$\dot{a}_i = (s_i - a_i b_i) b_i, \qquad \dot{b}_i = (s_i - a_i b_i) a_i.$$

Moreover,

$$\frac{\mathrm{d}}{\mathrm{d}t}(a_i^2 - b_i^2) = 0 \qquad \text{for every } i.$$

In particular, if the initialization is balanced, meaning $a_i(0) = b_i(0) > 0$ for every i, then $a_i(t) = b_i(t)$ for all t, and the amplitudes

$$\sigma_i(t) := a_i(t) b_i(t) = a_i(t)^2$$

obey the logistic equations

$$\dot{\sigma}_i = 2\sigma_i(s_i - \sigma_i).$$

Their explicit solutions are

$$\sigma_i(t) = \frac{s_i}{1 + \left(\frac{s_i}{\sigma_i(0)} - 1\right) e^{-2 s_i t}}.$$

Consequently, for any threshold $\theta \in (0, s_i)$, the hitting time $\tau_i(\theta)$ defined by $\sigma_i(\tau_i(\theta)) = \theta$ is

$$\tau_i(\theta) = \frac{1}{2 s_i} \log \left(\frac{\theta\,(s_i - \sigma_i(0))}{\sigma_i(0)(s_i - \theta)} \right).$$

For a fixed initial amplitude $\sigma_0 \in (0, \theta)$ and fixed threshold θ, the function

$$\tau(s; \sigma_0, \theta) := \frac{1}{2s} \log \left(\frac{\theta(s - \sigma_0)}{\sigma_0(s - \theta)} \right)$$

is strictly decreasing for $s > \theta$. Hence, if in addition $\sigma_i(0) = \sigma_0$ for all modes and $\theta < \min_i s_i$, then modes with larger singular values hit the threshold earlier.

Proof. Assume that for some time t the factors have the form

$$A = \sum_{i=1}^{r} a_i u_i e_i^\top, \qquad B = \sum_{i=1}^{r} b_i e_i v_i^\top.$$

Then

$$\begin{aligned} AB &= \left(\sum_{i=1}^{r} a_i u_i e_i^\top \right) \left(\sum_{j=1}^{r} b_j e_j v_j^\top \right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} a_i b_j u_i (e_i^\top e_j) v_j^\top \\ &= \sum_{i=1}^{r} a_i b_i u_i v_i^\top, \end{aligned}$$

using $e_i^\top e_j = \delta_{ij}$. Therefore

$$AB - T = \sum_{i=1}^{r} (a_i b_i - s_i) u_i v_i^\top.$$

To compute the gradient flow, first compute $B^\top$:

$$B^\top = \sum_{j=1}^{r} b_j v_j e_j^\top.$$

Hence

$$\begin{aligned} (AB - T)B^\top &= \left(\sum_{i=1}^{r} (a_i b_i - s_i) u_i v_i^\top \right) \left(\sum_{j=1}^{r} b_j v_j e_j^\top \right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{r} (a_i b_i - s_i) b_j u_i (v_i^\top v_j) e_j^\top \\ &= \sum_{i=1}^{r} (a_i b_i - s_i) b_i u_i e_i^\top, \end{aligned}$$

using $v_i^\top v_j = \delta_{ij}$. Therefore

$$\dot{A} = -(AB - T)B^\top = \sum_{i=1}^{r} (s_i - a_i b_i) b_i u_i e_i^\top.$$

This shows that the aligned form of A is preserved by the flow and that the coefficients obey

$$\dot{a}_i = (s_i - a_i b_i) b_i.$$

Now compute $A^\top$:

$$A^\top = \sum_{j=1}^{r} a_j e_j u_j^\top.$$

Then

$$\begin{aligned} A^\top(AB-T) &= \left(\sum_{j=1}^{r} a_j e_j u_j^\top\right)\left(\sum_{i=1}^{r}(a_i b_i - s_i)u_i v_i^\top\right) \\ &= \sum_{j=1}^{r}\sum_{i=1}^{r} a_j(a_i b_i - s_i)e_j(u_j^\top u_i)v_i^\top \\ &= \sum_{i=1}^{r} a_i(a_i b_i - s_i)e_i v_i^\top, \end{aligned}$$

using $u_j^\top u_i = \delta_{ji}$. Hence

$$\dot{B} = -A^\top(AB-T) = \sum_{i=1}^{r}(s_i - a_i b_i)a_i e_i v_i^\top.$$

This shows that the aligned form of B is also preserved and that

$$\dot{b}_i = (s_i - a_i b_i)a_i.$$

So the decoupled scalar ODEs are proved.

The conservation law follows by direct differentiation:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(a_i^2 - b_i^2) &= 2a_i\dot{a}_i - 2b_i\dot{b}_i \\ &= 2a_i(s_i - a_i b_i)b_i - 2b_i(s_i - a_i b_i)a_i \\ &= 0. \end{aligned}$$

Thus $a_i^2 - b_i^2$ is constant in time for every i.

If $a_i(0) = b_i(0)$, then this constant is zero, so

$$a_i(t)^2 = b_i(t)^2 \qquad \text{for all } t.$$

Because the coefficients are continuous in time and start positive, they remain on the positive branch as long as the solution exists, hence

$$a_i(t) = b_i(t).$$

In the balanced case, the mode amplitudes therefore satisfy a logistic equation.

Assume now that $a_i = b_i$ for all t. Define

$$\sigma_i := a_i b_i = a_i^2.$$

Then

$$\begin{aligned} \dot{\sigma}_i &= \frac{\mathrm{d}}{\mathrm{d}t}(a_i b_i) \\ &= \dot{a}_i b_i + a_i \dot{b}_i \\ &= (s_i - a_i b_i)b_i^2 + (s_i - a_i b_i)a_i^2 \\ &= (s_i - \sigma_i)(\sigma_i + \sigma_i) \\ &= 2\sigma_i(s_i - \sigma_i). \end{aligned}$$

This is the logistic ODE.

That logistic equation can be solved in closed form.

Fix one index i and write $\sigma = \sigma_i$, $s = s_i$ to simplify notation. The ODE is

$$\dot{\sigma} = 2\sigma(s - \sigma).$$

Assume $0 < \sigma(0) < s$, which holds when the initialization is small and positive. Separate variables:

$$\frac{\mathrm{d}\sigma}{\sigma(s-\sigma)} = 2\,\mathrm{d}t.$$

Use the partial-fraction decomposition

$$\frac{1}{\sigma(s-\sigma)} = \frac{1}{s}\left(\frac{1}{\sigma} + \frac{1}{s-\sigma}\right).$$

Hence

$$\int \frac{\mathrm{d}\sigma}{\sigma(s-\sigma)} = \frac{1}{s}\int \left(\frac{1}{\sigma} + \frac{1}{s-\sigma}\right)\mathrm{d}\sigma.$$

Because

$$\int \frac{1}{s-\sigma}\,\mathrm{d}\sigma = -\log(s-\sigma),$$

we obtain

$$\frac{1}{s}\big(\log\sigma - \log(s-\sigma)\big) = 2t + C.$$

Exponentiating yields

$$\frac{\sigma}{s-\sigma} = Ke^{2st}$$

for some constant $K > 0$. Solve for σ:

$$\sigma = (s-\sigma)Ke^{2st}$$

so

$$\sigma(1 + Ke^{2st}) = sKe^{2st}.$$

Therefore

$$\sigma(t) = \frac{sKe^{2st}}{1+Ke^{2st}} = \frac{s}{1+K^{-1}e^{-2st}}.$$

Use the initial condition $\sigma(0)$ to find K:

$$\sigma(0) = \frac{s}{1+K^{-1}} \quad \Longrightarrow \quad K^{-1} = \frac{s}{\sigma(0)} - 1.$$

Substitute this back into the formula. We obtain

$$\sigma(t) = \frac{s}{1 + \left(\frac{s}{\sigma(0)} - 1\right)e^{-2st}}.$$

This is the claimed explicit solution.

Solving the explicit formula for the hitting time gives the remaining claim. Fix a threshold $\theta \in (0, s_i)$ and solve $\sigma_i(t) = \theta$:

$$\theta = \frac{s_i}{1 + \left(\frac{s_i}{\sigma_i(0)} - 1\right)e^{-2s_i t}}.$$

Rearranging the equation gives

$$\begin{aligned}
\frac{s_i}{\theta} - 1 &= \left(\frac{s_i}{\sigma_i(0)} - 1\right)e^{-2s_i t}, \\
\frac{s_i - \theta}{\theta} &= \frac{s_i - \sigma_i(0)}{\sigma_i(0)}e^{-2s_i t}, \\
e^{-2s_i t} &= \frac{\sigma_i(0)(s_i - \theta)}{\theta(s_i - \sigma_i(0))}.
\end{aligned}$$

Taking logarithms gives

$$t = \frac{1}{2s_i}\log\left(\frac{\theta(s_i - \sigma_i(0))}{\sigma_i(0)(s_i - \theta)}\right).$$

This is the claimed expression for $\tau_i(\theta)$.

It remains to prove the monotonicity assertion under equal initial amplitudes. Fix $\sigma_0 \in (0, \theta)$ and define

$$\tau(s; \sigma_0, \theta) := \frac{1}{2s} \log\left(\frac{\theta(s-\sigma_0)}{\sigma_0(s-\theta)}\right), \qquad s > \theta.$$

Differentiating gives

$$\frac{\mathrm{d}\tau}{\mathrm{d}s} = -\frac{1}{2s^2} \log\left(\frac{\theta(s-\sigma_0)}{\sigma_0(s-\theta)}\right) + \frac{1}{2s}\left(\frac{1}{s-\sigma_0} - \frac{1}{s-\theta}\right).$$

Because $\theta > \sigma_0$, the logarithm's argument is strictly larger than 1, so the first term is strictly negative. Also $s-\sigma_0 > s-\theta > 0$, so $(s-\sigma_0)^{-1} < (s-\theta)^{-1}$ and the second term is strictly negative. Therefore $\mathrm{d}\tau/\mathrm{d}s < 0$ for $s > \theta$. Consequently, if $\sigma_i(0) = \sigma_0$ for all modes and $\theta < \min_i s_i$, then larger singular values have smaller hitting times. Without the equal-amplitude initialization condition across modes, the hitting time also depends on $\sigma_i(0)$, so singular values alone need not determine the order. □

Remark I.2 (Interpretation). Under aligned balanced factorized gradient flow with equal-amplitude initialization, larger singular modes of the surrogate reach fixed amplitude thresholds earlier. Combined with Theorem E.1, this provides a stylized mechanism for the empirical emergence ordering: cross-entropy defines a low-rank target T, and in this idealized factorized dynamics, equally initialized singular modes are amplified in decreasing order of strength, so interaction directions that are simultaneously common and useful reach any fixed linear-accessibility threshold first.

J. Scope of the pre-linearization claim

This section formalizes the objective-level meaning of "pre-linearization" used here. We fix a dictionary, restrict to width-k students, and compare bounded-logit cross-entropy minimizers with the rank-k surrogate subspace. Under the stated curvature and gap assumptions, the two selected subspaces are close in Frobenius projector distance.

Corollary J.1 (Objective-level induction of pre-linearization)**.** *Assume the exact-dictionary regime $r = 0$, so that*

$$\eta_* = C\phi.$$

Fix $k < q$, where $q = \mathrm{rank}(T)$ and

$$T = C\Sigma^{1/2} = U\,\mathrm{diag}(s_1, \ldots, s_q, 0, \ldots, 0)V^\top,$$

with $s_1 \geq s_2 \geq \cdots \geq s_q > 0$. Let $M_{\mathrm{CE},k}$ be any rank-at-most-k minimizer of the true cross-entropy risk over the admissible bounded-logit class, and let M_k be the surrogate-optimal width-k student from Theorem F.1. Assume that M_k is also admissible in the same bounded-logit class, i.e. $M_k\phi(x) \in K_B$ for μ-almost every x. Then

$$\mathcal{E}(M_{\mathrm{CE},k}) \leq \kappa_B\, \mathcal{E}(M_k) = \kappa_B \sum_{i=k+1}^{q} s_i^2,$$

where

$$\mathcal{E}(M) := \left\|(C-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

If, in addition, the gap hypothesis of Theorem H.2 holds, then the row space of $M_{\mathrm{CE},k}\Sigma^{1/2}$ is close to the top-k right singular subspace of T, quantitatively:

$$\|P_{\mathrm{CE}} - P_*\|_{\mathrm{F}}^2 \leq \frac{2(\kappa_B - 1)E_*}{g},$$

$$E_* := \sum_{i=k+1}^{q} s_i^2, \quad g := s_k^2 - s_{k+1}^2.$$

In particular, when κ_B is close to 1 and the spectral gap is large relative to $(\kappa_B - 1)E_*$, the true cross-entropy optimum and the surrogate optimum have small projector distance. In this precise bottleneck sense, cross-entropy induces pre-linearization.

Proof. Because $r = 0$, Theorem E.1 says that for every admissible rank-at-most-k student M,

$$\tfrac{m_B}{2}\mathcal{E}(M) \le \mathcal{R}(M) - \mathcal{R}_* \le \tfrac{M_B}{2}\mathcal{E}(M).$$

Apply the lower bound to $M_{\mathrm{CE},k}$ and the upper bound to M_k. Since $M_{\mathrm{CE},k}$ minimizes the true risk over the admissible rank-at-most-k class, we have

$$\mathcal{R}(M_{\mathrm{CE},k}) \le \mathcal{R}(M_k).$$

Therefore

$$\frac{m_B}{2}\mathcal{E}(M_{\mathrm{CE},k}) \le \mathcal{R}(M_{\mathrm{CE},k}) - \mathcal{R}_* \le \mathcal{R}(M_k) - \mathcal{R}_* \le \frac{M_B}{2}\mathcal{E}(M_k).$$

Multiply through by $2/m_B$. This gives

$$\mathcal{E}(M_{\mathrm{CE},k}) \le \frac{M_B}{m_B}\mathcal{E}(M_k) = \kappa_B\mathcal{E}(M_k).$$

Now invoke Theorem F.1, which proves that

$$\mathcal{E}(M_k) = \sum_{i=k+1}^{q} s_i^2.$$

Substituting this identity yields

$$\mathcal{E}(M_{\mathrm{CE},k}) \le \kappa_B \sum_{i=k+1}^{q} s_i^2.$$

This is the first displayed claim.

For the second claim, we invoke Theorem H.2. The hypotheses are satisfied: we are in the exact-dictionary regime $r = 0$, $M_{\mathrm{CE},k}$ is a true cross-entropy minimizer over admissible rank-at-most-k matrices, M_k is admissible, and the stated gap condition is assumed. Therefore

$$\|P_{\mathrm{CE}} - P_*\|_{\mathrm{F}}^2 \le \frac{2(\kappa_B - 1)E_*}{g}.$$

Together, these estimates show that, when the bounded-logit distortion is mild and the spectral gap is sufficiently large, the width-k cross-entropy optimum lies close to the surrogate's low-rank interaction subspace. □

Remark J.2 (Scope of Corollary J.1). Corollary J.1 shows that cross-entropy induces pre-linearization at the level of the bottleneck objective. Under the stated gap condition, the true cross-entropy optimum selects the same interaction subspace as the surrogate. This does not imply that arbitrary deep-network optimization globally realizes this optimum; it establishes the narrower objective-level claim.

K. Middle layers: pulling the objective back through the downstream tail

The theory extends from the final logits back to an arbitrary middle layer. The relevant object is the downstream tail map from the chosen layer to the logits. If that tail is affine, the extension is exact. If that tail is only close to affine on the relevant task subspace, the extension is approximate with an explicit error term.

K.1. Setup for a fixed middle layer

Fix a layer index ℓ. We write the hidden state at that layer as a vector in $\mathbb{R}^d$. Let

$$F_\ell : U \to G$$

be the downstream tail map from layer ℓ to the zero-sum logits, defined on a convex set $U \subset \mathbb{R}^d$ that contains the teacher and student layer-ℓ states under consideration. Thus the teacher next-token distribution at this layer is

$$p_*(\cdot \mid x) = \mathrm{softmax}(F_\ell(h_*(x))).$$

We again choose an interaction dictionary

$$\phi = (\phi_1, \dots, \phi_m) : X \to \mathbb{R}^m, \qquad \Sigma = \mathbb{E}[\phi(X)\phi(X)^\top] \succ 0,$$

and we model the teacher layer-ℓ state by

$$h_*(x) = B\phi(x), \qquad B \in \mathbb{R}^{d\times m}.$$

A width-k student at layer ℓ is any state map of the form

$$h_M(x) = M\phi(x), \qquad M \in \mathbb{R}^{d\times m}, \qquad \operatorname{rank}(M) \le k.$$

The induced logits are

$$\eta_*(x) = F_\ell(B\phi(x)), \qquad \widehat{\eta}_M(x) = F_\ell(M\phi(x)).$$

The population cross-entropy risk of M at layer ℓ is

$$\mathcal{R}_\ell(M) := \mathbb{E}_{X,Y\sim p_*}\big[-\log \operatorname{softmax}(F_\ell(M\phi(X)))_Y\big],$$

and the corresponding Bayes risk is

$$\mathcal{R}_{\ell,*} := \mathbb{E}_{X,Y\sim p_*}\big[-\log p_*(Y \mid X)\big].$$

K.2. Exact theorem when the downstream tail is affine

Theorem K.1 (Exact middle-layer reduction for an affine downstream tail)**.** *Assume that the downstream tail from layer ℓ to the logits is affine:*

$$F_\ell(u) = J_\ell u + b_\ell, \qquad J_\ell \in \mathbb{R}^{n\times d}, \qquad b_\ell \in G.$$

Assume also that for every admissible student M under consideration,

$$F_\ell(B\phi(x)) \in K_B \qquad \textit{and} \qquad F_\ell(M\phi(x)) \in K_B$$

for μ-almost every x.

Then for every admissible M with $\operatorname{rank}(M) \le k$,

$$\begin{aligned}\tfrac{m_B}{2}\left\|J_\ell(B-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2 &\le \mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*}\\ &\le \tfrac{M_B}{2}\left\|J_\ell(B-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.\end{aligned}$$

If

$$T_\ell := J_\ell B\Sigma^{1/2}$$

has singular values

$$s_1^{(\ell)} \ge s_2^{(\ell)} \ge \cdots \ge 0,$$

then the induced surrogate objective is minimized by the rank-k truncated singular subspace of T_ℓ. In particular, the layer-ℓ interactions selected for pre-linearization are the top singular directions of T_ℓ.

If, in addition, the orthogonal-coordinate regime holds with columns

$$B = [b_1 \ \cdots \ b_m]$$

and dictionary variances $\nu_r = \mathbb{E}[\phi_r(X)^2]$, then the score of feature r at layer ℓ is

$$d_{r,\ell} = \nu_r \ \|J_\ell b_r\|_2^2.$$

Proof. Because the tail is affine,

$$\eta_*(x) = F_\ell(B\phi(x)) = J_\ell B\phi(x) + b_\ell,$$

and similarly

$$\widehat{\eta}_M(x) = F_\ell(M\phi(x)) = J_\ell M\phi(x) + b_\ell.$$

In particular, the student and teacher differ only through the coefficient matrices $J_\ell B$ and $J_\ell M$.

Conditioning on the context again expresses the excess risk as an expected KL divergence.

By the definition of $\mathcal{R}_\ell(M)$ and $\mathcal{R}_{\ell,*}$, the same conditioning argument used in the proof of Theorem E.1 gives

$$\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} = \mathbb{E}_X\Big[\mathrm{KL}\left(\mathrm{softmax}(\eta_*(X)) \,\|\, \mathrm{softmax}(\widehat{\eta}_M(X))\right)\Big].$$

Thus the problem has been reduced to comparing the teacher and student logit maps.

Applying Lemma D.3 pointwise gives, for any two logit vectors $u, v \in G$,

$$\mathrm{KL}\left(\mathrm{softmax}(v) \,\|\, \mathrm{softmax}(u)\right) = D_A(u, v).$$

Applying this pointwise with

$$v = \eta_*(X) \qquad \text{and} \qquad u = \widehat{\eta}_M(X)$$

shows that

$$\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} = \mathbb{E}_X\Big[D_A\big(\widehat{\eta}_M(X), \eta_*(X)\big)\Big].$$

By assumption, both $\eta_*(x)$ and $\widehat{\eta}_M(x)$ lie in the bounded logit box K_B for almost every x. Since K_B is convex, the line segment between them also lies in K_B. Therefore Lemma D.4 together with Proposition D.5 yields

$$\begin{aligned}\tfrac{m_B}{2}\,\|\widehat{\eta}_M(x) - \eta_*(x)\|_2^2 &\le D_A\big(\widehat{\eta}_M(x), \eta_*(x)\big)\\ &\le \tfrac{M_B}{2}\,\|\widehat{\eta}_M(x) - \eta_*(x)\|_2^2\,.\end{aligned}$$

Take expectation over X. We obtain

$$\begin{aligned}\tfrac{m_B}{2}\,\|\widehat{\eta}_M - \eta_*\|_{L_2(\mu;G)}^2 &\le \mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*}\\ &\le \tfrac{M_B}{2}\,\|\widehat{\eta}_M - \eta_*\|_{L_2(\mu;G)}^2\,.\end{aligned}$$

It remains to compute the $L_2(\mu; G)$ norm of the logit difference. Since teacher and student share the affine offset b_ℓ,

$$\begin{aligned}\widehat{\eta}_M(x) - \eta_*(x) &= \big(J_\ell M\phi(x) + b_\ell\big) - \big(J_\ell B\phi(x) + b_\ell\big)\\ &= J_\ell(M - B)\phi(x).\end{aligned}$$

Therefore

$$\|\widehat{\eta}_M - \eta_*\|_{L_2(\mu;G)}^2 = \mathbb{E}\,\|J_\ell(M - B)\phi(X)\|_2^2\,.$$

Expanding this norm exactly as in the proof of Theorem E.1, we find

$$\begin{aligned}&\mathbb{E}\,\|J_\ell(M - B)\phi(X)\|_2^2\\ &= \mathbb{E}\Big[\,\mathrm{tr}\left(J_\ell(M - B)\phi(X)\phi(X)^\top(M - B)^\top J_\ell^\top\right)\Big]\\ &= \mathrm{tr}\left(J_\ell(M - B)\,\mathbb{E}[\phi(X)\phi(X)^\top]\,(M - B)^\top J_\ell^\top\right)\\ &= \mathrm{tr}\left(J_\ell(M - B)\Sigma(M - B)^\top J_\ell^\top\right)\\ &= \left\|J_\ell(B - M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.\end{aligned}$$

Substituting this identity into the bounded-logit comparison yields

$$\begin{aligned}\tfrac{m_B}{2}\left\|J_\ell(B - M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2 &\le \mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*}\\ &\le \tfrac{M_B}{2}\left\|J_\ell(B - M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.\end{aligned}$$

This proves the first displayed inequality.

Define

$$C_\ell := J_\ell B.$$

Then the previous display shows that the induced surrogate is

$$\left\| T_\ell - J_\ell M \Sigma^{1/2} \right\|_{\mathrm{F}}^2, \qquad T_\ell := C_\ell \Sigma^{1/2} = J_\ell B \Sigma^{1/2},$$

with $\mathrm{rank}(M) \le k$. It remains to verify that the rank-k truncated SVD of T_ℓ is attainable within the admissible layer-ℓ class, rather than only among arbitrary rank-k logit-space matrices.

Let

$$T_{\ell,k} = U_k S_k V_k^\top$$

be a rank-k truncated SVD of T_ℓ. Since $T_\ell = J_\ell B \Sigma^{1/2}$, we have $\mathrm{col}(T_\ell) \subseteq \mathrm{col}(J_\ell)$. Hence every left singular vector of T_ℓ with nonzero singular value lies in $\mathrm{col}(J_\ell)$, so there exists $A_k \in \mathbb{R}^{d \times k}$ such that

$$J_\ell A_k = U_k S_k.$$

For instance, one may take $A_k = J_\ell^+ U_k S_k$, where J_ℓ^+ is the Moore–Penrose pseudoinverse. Define

$$M_k := A_k V_k^\top \Sigma^{-1/2}.$$

Then $\mathrm{rank}(M_k) \le k$ and

$$J_\ell M_k \Sigma^{1/2} = J_\ell A_k V_k^\top = U_k S_k V_k^\top = T_{\ell,k}.$$

Thus the Eckart–Young lower bound is attainable by an admissible layer-ℓ coefficient matrix M_k. It follows that the optimal width-k layer-ℓ representation is obtained by keeping a top-k right singular subspace of T_ℓ.

Suppose now that the orthogonal-coordinate regime holds. Then the teacher layer-ℓ logit map can be written as

$$\eta_*(x) - b_\ell = J_\ell B \phi(x) = \sum_{r=1}^{m} (J_\ell b_r)\, \phi_r(x).$$

Thus the output direction associated with feature r is exactly

$$c_{r,\ell} := J_\ell b_r.$$

Theorem G.2 has already proved that in the orthogonal-coordinate regime the feature score is

$$d_{r,\ell} = \nu_r \, \|c_{r,\ell}\|_2^2 \,.$$

Substituting $c_{r,\ell} = J_\ell b_r$ gives

$$d_{r,\ell} = \nu_r \, \|J_\ell b_r\|_2^2 \,.$$

This is the claimed formula. □

Remark K.2 (Interpretation of the middle-layer affine theorem). Theorem K.1 says that pre-linearization at a middle layer is governed not by the final readout alone but by the final readout *pulled back through the downstream tail*. A local interaction direction b_r receives score $\nu_r \, \|J_\ell b_r\|_2^2$. The factor ν_r measures how much the interaction occurs in the data; the factor $\|J_\ell b_r\|_2^2$ measures how strongly that direction would affect the logits if it were made linearly available in the layer-ℓ state.

K.3. Approximate theorem when the downstream tail is only locally linear

The exact affine theorem covers the special case of an exactly affine downstream tail; in trained networks the tail is generally nonlinear. The next theorem requires only that the tail be *close* to affine on the relevant task subspace.

Theorem K.3 (Middle-layer reduction for a locally linear downstream tail)**.** *Let $U \subset \mathbb{R}^d$ be a convex set, and let*

$$F_\ell : U \to G$$

be a C^1 map. Let $S \subset \mathbb{R}^d$ be a linear subspace. Assume the teacher and all admissible students have the form

$$h_*(x) = B\phi(x), \qquad h_M(x) = M\phi(x),$$

with

$$\mathrm{range}(B) \subset S, \qquad \mathrm{range}(M) \subset S, \qquad \mathrm{rank}(M) \le k,$$

and assume that

$$h_*(x) \in U, \quad h_M(x) \in U,$$
$$F_\ell(h_*(x)) \in K_B, \quad F_\ell(h_M(x)) \in K_B$$

for μ-almost every x.

Fix a matrix $J_\ell \in \mathbb{R}^{n\times d}$ and constants $\alpha_\ell > 0$, $\delta_\ell \ge 0$ such that

$$\sup_{u\in U} \|DF_\ell(u) - J_\ell\|_{\mathrm{op}} \le \delta_\ell$$

and

$$\|J_\ell v\|_2 \ge \alpha_\ell \|v\|_2 \qquad \textit{for every } v \in S.$$

Assume moreover that $\delta_\ell < \alpha_\ell$.

Then every admissible student M satisfies

$$\begin{aligned}
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&\ge \tfrac{m_B}{2}\Big(1 - \tfrac{\delta_\ell}{\alpha_\ell}\Big)^2 \Big\| J_\ell(B-M)\Sigma^{1/2}\Big\|_{\mathrm{F}}^2, \\
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&\le \tfrac{M_B}{2}\Big(1 + \tfrac{\delta_\ell}{\alpha_\ell}\Big)^2 \Big\| J_\ell(B-M)\Sigma^{1/2}\Big\|_{\mathrm{F}}^2.
\end{aligned}$$

Consequently, the true middle-layer cross-entropy objective is uniformly comparable to the same rank-k surrogate as in Theorem K.1, but with multiplicative distortion factor

$$\kappa_B \left(\frac{1+\delta_\ell/\alpha_\ell}{1-\delta_\ell/\alpha_\ell}\right)^2.$$

Proof. Exactly as before,

$$\begin{aligned}
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&= \mathbb{E}_X\Big[\mathrm{KL}\left(\mathrm{softmax}(F_\ell(h_*(X)))\,\|\,\mathrm{softmax}(F_\ell(h_M(X)))\right)\Big].
\end{aligned}$$

By Lemma D.3, this is

$$\begin{aligned}
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&= \mathbb{E}_X\Big[D_A\big(F_\ell(h_M(X)), F_\ell(h_*(X))\big)\Big].
\end{aligned}$$

For almost every x, both $F_\ell(h_*(x))$ and $F_\ell(h_M(x))$ lie in K_B, so Proposition D.5 gives

$$\begin{aligned}
&D_A\big(F_\ell(h_M(x)), F_\ell(h_*(x))\big) \\
&\ge \tfrac{m_B}{2}\|F_\ell(h_*(x)) - F_\ell(h_M(x))\|_2^2, \\
&D_A\big(F_\ell(h_M(x)), F_\ell(h_*(x))\big) \\
&\le \tfrac{M_B}{2}\|F_\ell(h_*(x)) - F_\ell(h_M(x))\|_2^2.
\end{aligned}$$

Take expectation over X. We obtain

$$\begin{aligned}
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&\ge \tfrac{m_B}{2}\|F_\ell(h_*) - F_\ell(h_M)\|_{L_2(\mu;G)}^2, \\
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&\le \tfrac{M_B}{2}\|F_\ell(h_*) - F_\ell(h_M)\|_{L_2(\mu;G)}^2.
\end{aligned}$$

It remains to compare the nonlinear logit difference with its linearized approximation, using the fundamental theorem of calculus along the teacher-student segment.

Fix a context x. Define

$$\begin{aligned} u &:= h_M(x) = M\phi(x), \\ v &:= h_*(x) = B\phi(x), \\ d &:= v - u = (B - M)\phi(x). \end{aligned}$$

Because both v and u lie in U and U is convex, the line segment

$$u + td, \qquad 0 \le t \le 1,$$

remains inside U. Since F_ℓ is C^1 on U, the fundamental theorem of calculus along this segment gives

$$F_\ell(v) - F_\ell(u) = \int_0^1 DF_\ell(u + td)\, d \, \mathrm{d}t.$$

Subtract the linear term $J_\ell d$ from both sides:

$$F_\ell(v) - F_\ell(u) - J_\ell d = \int_0^1 \big(DF_\ell(u + td) - J_\ell\big) d \, \mathrm{d}t.$$

Take norms and use the operator-norm bound:

$$\begin{aligned} \|F_\ell(v) - F_\ell(u) - J_\ell d\|_2 &\le \int_0^1 \|DF_\ell(u + td) - J_\ell\|_{\mathrm{op}} \, \|d\|_2 \, \mathrm{d}t \\ &\le \int_0^1 \delta_\ell \, \|d\|_2 \, \mathrm{d}t \\ &= \delta_\ell \, \|d\|_2 \, . \end{aligned}$$

This is the basic local-linearity remainder bound.

The lower singular-value bound for J_ℓ on the task subspace converts this absolute remainder estimate into a relative one.

Because both $\mathrm{range}(B)$ and $\mathrm{range}(M)$ lie inside S, the difference vector $d = (B - M)\phi(x)$ also lies in S. By the lower singular-value assumption on $J_\ell|_S$,

$$\|J_\ell d\|_2 \ge \alpha_\ell \, \|d\|_2 \, .$$

Rearrange this inequality:

$$\|d\|_2 \le \frac{1}{\alpha_\ell} \, \|J_\ell d\|_2 \, .$$

Substituting this into the remainder estimate gives

$$\|F_\ell(v) - F_\ell(u) - J_\ell d\|_2 \le \frac{\delta_\ell}{\alpha_\ell} \, \|J_\ell d\|_2 \, .$$

Now use the triangle inequality and the reverse triangle inequality. First,

$$\begin{aligned} \|F_\ell(v) - F_\ell(u)\|_2 &\le \|J_\ell d\|_2 + \|F_\ell(v) - F_\ell(u) - J_\ell d\|_2 \\ &\le \left(1 + \frac{\delta_\ell}{\alpha_\ell}\right) \|J_\ell d\|_2 \, . \end{aligned}$$

Second,

$$\begin{aligned} \|F_\ell(v) - F_\ell(u)\|_2 &\ge \|J_\ell d\|_2 - \|F_\ell(v) - F_\ell(u) - J_\ell d\|_2 \\ &\ge \left(1 - \frac{\delta_\ell}{\alpha_\ell}\right) \|J_\ell d\|_2 \, . \end{aligned}$$

Because $\delta_\ell < \alpha_\ell$, the lower multiplicative factor is strictly positive.

Squaring the pointwise bounds and averaging over the data distribution gives the desired Frobenius comparison.

The pointwise bounds above hold for almost every x. Squaring them gives

$$\begin{aligned}
&\|F_\ell(h_*(x)) - F_\ell(h_M(x))\|_2^2 \\
&\geq \left(1 - \tfrac{\delta_\ell}{\alpha_\ell}\right)^2 \|J_\ell(B-M)\phi(x)\|_2^2\,, \\
&\|F_\ell(h_*(x)) - F_\ell(h_M(x))\|_2^2 \\
&\leq \left(1 + \tfrac{\delta_\ell}{\alpha_\ell}\right)^2 \|J_\ell(B-M)\phi(x)\|_2^2\,.
\end{aligned}$$

Take expectation over X:

$$\begin{aligned}
&\|F_\ell(h_*) - F_\ell(h_M)\|_{L_2(\mu;G)}^2 \\
&\geq \left(1 - \tfrac{\delta_\ell}{\alpha_\ell}\right)^2 \mathbb{E}\,\|J_\ell(B-M)\phi(X)\|_2^2\,, \\
&\|F_\ell(h_*) - F_\ell(h_M)\|_{L_2(\mu;G)}^2 \\
&\leq \left(1 + \tfrac{\delta_\ell}{\alpha_\ell}\right)^2 \mathbb{E}\,\|J_\ell(B-M)\phi(X)\|_2^2\,.
\end{aligned}$$

The common expectation equals, as above,

$$\mathbb{E}\,\|J_\ell(B-M)\phi(X)\|_2^2 = \left\|J_\ell(B-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

Substituting this identity back into the previous display, and then into the bounded-logit sandwich, yields

$$\begin{aligned}
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&\geq \tfrac{m_B}{2}\left(1 - \tfrac{\delta_\ell}{\alpha_\ell}\right)^2 \left\|J_\ell(B-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2, \\
&\mathcal{R}_\ell(M) - \mathcal{R}_{\ell,*} \\
&\leq \tfrac{M_B}{2}\left(1 + \tfrac{\delta_\ell}{\alpha_\ell}\right)^2 \left\|J_\ell(B-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.
\end{aligned}$$

This is exactly the claimed inequality.

The previous display shows that the true risk is squeezed between two constant multiples of the same surrogate

$$\left\|J_\ell(B-M)\Sigma^{1/2}\right\|_{\mathrm{F}}^2.$$

The ratio between the upper and lower multiplicative constants is

$$\frac{M_B\,(1+\delta_\ell/\alpha_\ell)^2}{m_B\,(1-\delta_\ell/\alpha_\ell)^2} = \kappa_B\left(\frac{1+\delta_\ell/\alpha_\ell}{1-\delta_\ell/\alpha_\ell}\right)^2.$$

which proves the stated distortion factor. □

Remark K.4 (Interpretation). Theorem K.3 provides the main reduction for middle layers. If the downstream tail is close to a fixed linear map J_ℓ on the relevant task subspace, then the layer-ℓ cross-entropy objective is comparable to the affine-case rank-k surrogate, with an explicit distortion factor measuring the tail's nonlinearity. The effective operator remains

$$T_\ell = J_\ell B \Sigma^{1/2},$$

so the dominant pre-linearized directions are the dominant singular directions of the interaction operator pulled back by the downstream Jacobian.

K.4. Residual-stream specializations

We specialize the middle-layer theorem to the standard residual-stream formulation of transformers.

Corollary K.5 (Exact residual-stream theorem for affine residual blocks)**.** *Fix a layer index ℓ. Suppose the residual stream from layer ℓ onward evolves by affine residual updates*

$$h_{t+1} = h_t + A_t h_t + a_t = (I + A_t)h_t + a_t, \qquad t = \ell, \dots, L-1,$$

where each $A_t \in \mathbb{R}^{d\times d}$ and $a_t \in \mathbb{R}^d$. Suppose the final readout is affine:

$$\eta = W h_L + b, \qquad W \in \mathbb{R}^{n\times d}, \qquad b \in G.$$

Let the teacher and student layer-ℓ states be $B\phi(x)$ and $M\phi(x)$, respectively.

Then there exist a matrix $\Pi_\ell \in \mathbb{R}^{d\times d}$ and a vector $c_\ell \in \mathbb{R}^d$ such that

$$h_L = \Pi_\ell h_\ell + c_\ell, \qquad \Pi_\ell := (I + A_{L-1}) \cdots (I + A_\ell),$$

and consequently the full downstream tail from layer ℓ to the logits is affine:

$$F_\ell(u) = W\Pi_\ell u + (Wc_\ell + b).$$

Therefore Theorem K.1 applies with

$$J_\ell = W\Pi_\ell.$$

In the orthogonal-coordinate regime, the layer-ℓ score of feature r is

$$d_{r,\ell} = \nu_r \, \|W\Pi_\ell b_r\|_2^2 \,.$$

Proof. For every integer j with $\ell \le j \le L$, we claim that there exist a matrix $\Pi_{\ell:j}$ and a vector $c_{\ell:j}$ such that

$$h_j = \Pi_{\ell:j} h_\ell + c_{\ell:j}.$$

We prove this by induction on j.

When $j = \ell$, the claim is immediate with

$$\Pi_{\ell:\ell} = I, \qquad c_{\ell:\ell} = 0.$$

Now assume the claim has been proved for some j with $\ell \le j < L$. Then

$$h_j = \Pi_{\ell:j} h_\ell + c_{\ell:j}.$$

Apply the residual update at layer j:

$$\begin{aligned} h_{j+1} &= (I + A_j)h_j + a_j \\ &= (I + A_j)(\Pi_{\ell:j} h_\ell + c_{\ell:j}) + a_j \\ &= \big((I + A_j)\Pi_{\ell:j}\big)h_\ell + \big((I + A_j)c_{\ell:j} + a_j\big). \end{aligned}$$

Therefore the claim also holds for $j + 1$, with

$$\Pi_{\ell:j+1} = (I + A_j)\Pi_{\ell:j}, \qquad c_{\ell:j+1} = (I + A_j)c_{\ell:j} + a_j.$$

By induction, the representation

$$h_j = \Pi_{\ell:j} h_\ell + c_{\ell:j}$$

holds for every j from ℓ to L.

The recursion identifies the explicit downstream matrix product.

The recurrence for $\Pi_{\ell:j}$ starts from the identity and multiplies by $(I + A_j)$ at each subsequent layer. Therefore

$$\Pi_{\ell:L} = (I + A_{L-1}) \cdots (I + A_\ell).$$

For notational simplicity, denote this matrix by

$$\Pi_\ell := \Pi_{\ell:L}.$$

Likewise, denote

$$c_\ell := c_{\ell:L}.$$

Then the full downstream relation is

$$h_L = \Pi_\ell h_\ell + c_\ell.$$

This proves the first displayed formula in the statement.

Substituting this unrolled form into the final readout gives the downstream logit tail explicitly.

The final logits are

$$\eta = W h_L + b.$$

Substituting the affine expression for h_L obtained above,

$$\begin{aligned}\eta &= W(\Pi_\ell h_\ell + c_\ell) + b \\ &= W\Pi_\ell h_\ell + (W c_\ell + b).\end{aligned}$$

Therefore the entire tail map from layer ℓ to the logits is the affine map

$$F_\ell(u) = W\Pi_\ell u + (W c_\ell + b).$$

So Theorem K.1 applies with

$$J_\ell = W\Pi_\ell.$$

The final claim about the feature score now follows directly from the orthogonal-coordinate part of Theorem K.1. Substitute $J_\ell = W\Pi_\ell$ into

$$d_{r,\ell} = \nu_r \, \|J_\ell b_r\|_2^2 \,.$$

This gives

$$d_{r,\ell} = \nu_r \, \|W\Pi_\ell b_r\|_2^2 \,.$$

That is exactly the announced residual-stream score. □

Lemma K.6 (Telescoping identity for differences of matrix products)**.** *Let $P_{L-1}, \dots, P_\ell$ and $Q_{L-1}, \dots, Q_\ell$ be matrices of compatible dimensions. Then*

$$\begin{aligned}&P_{L-1} \cdots P_\ell - Q_{L-1} \cdots Q_\ell \\ &= \sum_{s=\ell}^{L-1} P_{L-1} \cdots P_{s+1} \, (P_s - Q_s) \, Q_{s-1} \cdots Q_\ell,\end{aligned}$$

where an empty product is interpreted as the identity matrix.

Proof. We prove the identity by induction on the number of factors.

Base case: one factor.

If $L - 1 = \ell$, then the left-hand side is simply

$$P_\ell - Q_\ell.$$

The sum on the right-hand side has only one term, namely

$$(P_\ell - Q_\ell),$$

because both side products are empty and therefore equal to the identity. So the identity holds in the base case.

Inductive step.

Assume the identity holds whenever the number of factors is N. We prove it for $N+1$ factors. Write

$$\begin{aligned}&P_{L-1}\cdots P_\ell - Q_{L-1}\cdots Q_\ell\\&= P_{L-1}(P_{L-2}\cdots P_\ell) - Q_{L-1}(Q_{L-2}\cdots Q_\ell).\end{aligned}$$

Add and subtract the mixed term $P_{L-1}(Q_{L-2}\cdots Q_\ell)$:

$$\begin{aligned}P_{L-1}\cdots P_\ell - Q_{L-1}\cdots Q_\ell = P_{L-1}(P_{L-2}\cdots P_\ell - Q_{L-2}\cdots Q_\ell)\\+ (P_{L-1}-Q_{L-1})Q_{L-2}\cdots Q_\ell.\end{aligned}$$

By the induction hypothesis, the first bracket equals

$$\sum_{s=\ell}^{L-2} P_{L-2}\cdots P_{s+1}(P_s - Q_s)Q_{s-1}\cdots Q_\ell.$$

Multiply this whole sum on the left by P_{L-1}. We obtain

$$\sum_{s=\ell}^{L-2} P_{L-1}\cdots P_{s+1}(P_s - Q_s)Q_{s-1}\cdots Q_\ell.$$

Then add the remaining term

$$(P_{L-1}-Q_{L-1})Q_{L-2}\cdots Q_\ell,$$

which is exactly the $s = L-1$ term of the desired sum. Therefore

$$\begin{aligned}&P_{L-1}\cdots P_\ell - Q_{L-1}\cdots Q_\ell\\&= \sum_{s=\ell}^{L-1} P_{L-1}\cdots P_{s+1}(P_s - Q_s)Q_{s-1}\cdots Q_\ell.\end{aligned}$$

This completes the induction. □

Proposition K.7 (Approximate residual-stream Jacobian under locally linear residual blocks)**.** *Let*

$$\psi_t(u) := u + g_t(u), \qquad t = \ell, \dots, L-1,$$

where each $g_t : U \to \mathbb{R}^d$ is C^1 on a convex set $U \subset \mathbb{R}^d$, and assume $\psi_t(U) \subset U$ for all t. Let

$$\Psi_{\ell:L} := \psi_{L-1} \circ \cdots \circ \psi_\ell$$

be the downstream residual map. Suppose that for each t there exist a matrix $A_t \in \mathbb{R}^{d\times d}$ and a number $\epsilon_t \geq 0$ such that

$$\sup_{u\in U} \|Dg_t(u) - A_t\|_{\mathrm{op}} \leq \epsilon_t.$$

Define

$$Q_t := I + A_t, \qquad \Pi_\ell := Q_{L-1}\cdots Q_\ell.$$

Then for every $u \in U$,

$$\|D\Psi_{\ell:L}(u) - \Pi_\ell\|_{\mathrm{op}} \leq \Delta_\ell,$$

where

$$\Delta_\ell := \sum_{s=\ell}^{L-1}\left(\prod_{t=L-1}^{s+1}(\|Q_t\|_{\mathrm{op}} + \epsilon_t)\right)\epsilon_s\left(\prod_{t=s-1}^{\ell}\|Q_t\|_{\mathrm{op}}\right),$$

with empty products again interpreted as 1.

Consequently, if the final readout is affine,

$$F_\ell(u) = W_{\mathrm{out}}\,\Psi_{\ell:L}(u) + b, \qquad W_{\mathrm{out}} \in \mathbb{R}^{n\times d}, \qquad b \in G,$$

then

$$\sup_{u\in U} \|DF_\ell(u) - W_{\text{out}}\Pi_\ell\|_{\text{op}} \le \|W_{\text{out}}\|_{\text{op}}\ \Delta_\ell.$$

Therefore Theorem K.3 applies with

$$J_\ell = W_{\text{out}}\Pi_\ell, \qquad \delta_\ell = \|W_{\text{out}}\|_{\text{op}}\ \Delta_\ell,$$

provided the lower singular-value assumption on J_ℓ holds on the chosen interaction subspace.

Proof. Fix $u \in U$. Define the intermediate states recursively by

$$h_\ell(u) := u, \qquad h_{t+1}(u) := \psi_t(h_t(u)) \qquad (t = \ell, \dots, L-1).$$

Because each ψ_t maps U into itself, every intermediate state $h_t(u)$ remains in U.

Now define the matrices

$$P_t(u) := D\psi_t(h_t(u)) = I + Dg_t(h_t(u)), \qquad Q_t := I + A_t.$$

By the chain rule,

$$D\Psi_{\ell:L}(u) = P_{L-1}(u)\cdots P_\ell(u).$$

Therefore

$$D\Psi_{\ell:L}(u) - \Pi_\ell = P_{L-1}(u)\cdots P_\ell(u) - Q_{L-1}\cdots Q_\ell.$$

Lemma K.6 gives the exact identity

$$P_{L-1}(u)\cdots P_\ell(u) - Q_{L-1}\cdots Q_\ell = \sum_{s=\ell}^{L-1} P_{L-1}(u)\cdots P_{s+1}(u)\,(P_s(u) - Q_s)\,Q_{s-1}\cdots Q_\ell.$$

Take operator norms and use the triangle inequality:

$$\|D\Psi_{\ell:L}(u) - \Pi_\ell\|_{\text{op}} \le \sum_{s=\ell}^{L-1} \|P_{L-1}(u)\cdots P_{s+1}(u)\|_{\text{op}}\ \|P_s(u) - Q_s\|_{\text{op}}\ \|Q_{s-1}\cdots Q_\ell\|_{\text{op}}.$$

Each factor in the telescoping sum can be bounded using the local Jacobian error assumptions. By definition,

$$P_s(u) - Q_s = Dg_s(h_s(u)) - A_s.$$

Since $h_s(u) \in U$, the hypothesis implies

$$\|P_s(u) - Q_s\|_{\text{op}} = \|Dg_s(h_s(u)) - A_s\|_{\text{op}} \le \epsilon_s.$$

Also,

$$\|P_t(u)\|_{\text{op}} \le \|Q_t\|_{\text{op}} + \|P_t(u) - Q_t\|_{\text{op}} \le \|Q_t\|_{\text{op}} + \epsilon_t.$$

Finally,

$$\|Q_{s-1}\cdots Q_\ell\|_{\text{op}} \le \prod_{t=s-1}^{\ell} \|Q_t\|_{\text{op}}$$

by repeated submultiplicativity of the operator norm. Similarly,

$$\|P_{L-1}(u)\cdots P_{s+1}(u)\|_{\text{op}} \le \prod_{t=L-1}^{s+1} (\|Q_t\|_{\text{op}} + \epsilon_t).$$

Substituting these three bounds into the telescoping estimate gives

$$\begin{aligned}&\|D\Psi_{\ell:L}(u) - \Pi_\ell\|_{\mathrm{op}}\\ &\le \sum_{s=\ell}^{L-1}\left(\prod_{t=L-1}^{s+1}(\|Q_t\|_{\mathrm{op}} + \epsilon_t)\right)\epsilon_s\left(\prod_{t=s-1}^{\ell}\|Q_t\|_{\mathrm{op}}\right).\end{aligned}$$

The right-hand side is exactly Δ_ℓ. Therefore

$$\|D\Psi_{\ell:L}(u) - \Pi_\ell\|_{\mathrm{op}} \le \Delta_\ell.$$

This proves the first displayed bound.

If the final readout is

$$F_\ell(u) = W_{\mathrm{out}}\,\Psi_{\ell:L}(u) + b,$$

then differentiation gives

$$DF_\ell(u) = W_{\mathrm{out}}\,D\Psi_{\ell:L}(u).$$

Therefore

$$\begin{aligned}\|DF_\ell(u) - W_{\mathrm{out}}\Pi_\ell\|_{\mathrm{op}} &= \big\|W_{\mathrm{out}}\big(D\Psi_{\ell:L}(u) - \Pi_\ell\big)\big\|_{\mathrm{op}}\\ &\le \|W_{\mathrm{out}}\|_{\mathrm{op}}\,\|D\Psi_{\ell:L}(u) - \Pi_\ell\|_{\mathrm{op}}\\ &\le \|W_{\mathrm{out}}\|_{\mathrm{op}}\,\Delta_\ell.\end{aligned}$$

Since this holds for every $u \in U$, we obtain

$$\sup_{u\in U}\|DF_\ell(u) - W_{\mathrm{out}}\Pi_\ell\|_{\mathrm{op}} \le \|W_{\mathrm{out}}\|_{\mathrm{op}}\,\Delta_\ell.$$

So Theorem K.3 applies with

$$J_\ell = W_{\mathrm{out}}\Pi_\ell, \qquad \delta_\ell = \|W_{\mathrm{out}}\|_{\mathrm{op}}\,\Delta_\ell,$$

provided the stated lower singular-value assumption on J_ℓ holds on the chosen task subspace. □

Proposition K.8 (Residual updates with small Lipschitz constant preserve direction norms)**.** *Let*

$$\psi_t(u) = u + g_t(u)$$

with $\mathrm{Lip}(g_t) \le \rho_t < 1$. *Then for all* $u, v \in U$,

$$(1-\rho_t)\,\|u-v\|_2 \le \|\psi_t(u) - \psi_t(v)\|_2 \le (1+\rho_t)\,\|u-v\|_2\,.$$

Consequently, for the composition

$$\Psi_{\ell:j} := \psi_{j-1}\circ\cdots\circ\psi_\ell, \qquad \ell < j \le L,$$

one has

$$\begin{aligned}\left(\prod_{t=\ell}^{j-1}(1-\rho_t)\right)\|u-v\|_2 &\le \|\Psi_{\ell:j}(u) - \Psi_{\ell:j}(v)\|_2\\ &\le \left(\prod_{t=\ell}^{j-1}(1+\rho_t)\right)\|u-v\|_2\,.\end{aligned}$$

Proof. For a single residual block,

$$\psi_t(u) - \psi_t(v) = (u-v) + \big(g_t(u) - g_t(v)\big).$$

Take norms and use the triangle inequality:

$$\begin{aligned}\|\psi_t(u) - \psi_t(v)\|_2 &\le \|u - v\|_2 + \|g_t(u) - g_t(v)\|_2 \\ &\le (1 + \rho_t)\,\|u - v\|_2 .\end{aligned}$$

Now use the reverse triangle inequality:

$$\begin{aligned}\|\psi_t(u) - \psi_t(v)\|_2 &\ge \|u - v\|_2 - \|g_t(u) - g_t(v)\|_2 \\ &\ge (1 - \rho_t)\,\|u - v\|_2 .\end{aligned}$$

This proves the one-step bound.

For the multi-step bound, apply the one-step inequality repeatedly. For the upper bound,

$$\begin{aligned}&\|\Psi_{\ell:j}(u) - \Psi_{\ell:j}(v)\|_2 \\ &\le (1 + \rho_{j-1})\,\|\Psi_{\ell:j-1}(u) - \Psi_{\ell:j-1}(v)\|_2 \\ &\le (1 + \rho_{j-1})(1 + \rho_{j-2})\,\|\Psi_{\ell:j-2}(u) - \Psi_{\ell:j-2}(v)\|_2 \\ &\vdots \\ &\le \left(\prod_{t=\ell}^{j-1}(1 + \rho_t)\right)\|u - v\|_2 .\end{aligned}$$

The same iteration applied to the one-step lower bound yields the product with factors $(1 - \rho_t)$. This proves the proposition. □

Remark K.9 (Consequences for residual-stream representations). The residual stream matters for two distinct reasons. First, Corollary K.5 and Proposition K.7 show that the middle-layer pre-linearization score is obtained by pulling the readout back through the downstream residual tail. Second, Proposition K.8 shows that when each residual update is not too large relative to the identity skip, a residual-stream direction introduced at an earlier layer cannot be contracted below a computable fraction of its original norm by later layers. Together these two facts explain why middle-layer pre-linearization should be both useful for the final loss and observable in later residual-stream activations.

K.5. The same singular subspace is optimal for local control

Theorem K.10 (Top singular directions are optimal first-order steering directions in whitened coordinates)**.** *Fix a layer ℓ and suppose local logit changes in whitened feature coordinates are approximated by*

$$\Delta\eta \approx T_\ell u, \qquad u \in \mathbb{R}^m.$$

For any k-dimensional subspace $S \subseteq \mathbb{R}^m$ with orthogonal projector P_S, define the average first-order logit energy of isotropic unit perturbations supported on S by

$$\mathcal{C}_\ell(S) := \mathbb{E}_{u\sim\mathrm{Unif}(S\cap\mathbb{S}^{m-1})}\,\|T_\ell u\|_2^2 = \frac{1}{k}\,\mathrm{tr}(P_S T_\ell^\top T_\ell).$$

Then $\mathcal{C}_\ell(S)$ is maximized when S is the top-k right singular subspace of T_ℓ, and the optimum value is

$$\max_{\dim S = k} \mathcal{C}_\ell(S) = \frac{1}{k}\sum_{i=1}^{k} s_i(T_\ell)^2.$$

In particular, the top right singular vector is the most effective unit-norm first-order steering direction in whitened feature-coordinate norm.

Proof. If $V_S \in \mathbb{R}^{m\times k}$ has orthonormal columns spanning S, then $P_S = V_S V_S^\top$. The covariance of a uniformly random unit vector on S is P_S/k, so

$$\mathcal{C}_\ell(S) = \mathrm{tr}\left(T_\ell^\top T_\ell\,\frac{P_S}{k}\right) = \frac{1}{k}\,\mathrm{tr}(P_S T_\ell^\top T_\ell).$$

Thus maximizing $\mathcal{C}_\ell(S)$ over k-dimensional subspaces is equivalent to maximizing $\mathrm{tr}(P_S T_\ell^\top T_\ell)$. The Ky Fan projection argument used in the rank-k approximation theorem shows that the maximizing projector is onto the top-k eigenspace of $T_\ell^\top T_\ell$. Equivalently, the optimal subspace is the top-k right singular subspace of T_ℓ, and the optimum value is the average of the top k squared singular values. □

Theorem K.10 is the control-theoretic counterpart of Theorem E.1. The spectral directions that cross-entropy preserves through a bottleneck are also the directions that give the largest first-order downstream leverage per unit whitened-feature-coordinate norm.

Proposition K.11 (Hidden-state-norm local control)**.** *Let*

$$A_\ell := B\Sigma^{1/2}, \qquad G_\ell := A_\ell^\top A_\ell.$$

A whitened feature-coordinate perturbation u corresponds to the hidden-state perturbation $\delta h = A_\ell u$ and to first-order logit change $T_\ell u$. Assume that $G_\ell \succ 0$ on the feature-coordinate subspace under consideration. If the intervention budget is hidden Euclidean norm, then the one-dimensional objective is the generalized Rayleigh quotient

$$\frac{\|T_\ell u\|_2^2}{\|A_\ell u\|_2^2} = \frac{u^\top T_\ell^\top T_\ell u}{u^\top G_\ell u}.$$

Its stationary directions solve the generalized eigenproblem

$$T_\ell^\top T_\ell u = \lambda G_\ell u.$$

Equivalently, after the change of variables $w = G_\ell^{1/2} u$, the optimal k-dimensional hidden-norm control subspace is obtained from the top k right singular directions of

$$T_\ell G_\ell^{-1/2}.$$

The whitened-coordinate result of Theorem K.10 is recovered when $G_\ell = I$.

Proof. The hidden perturbation associated with u is $\delta h = A_\ell u$, so the hidden-state norm constraint is

$$\|\delta h\|_2^2 = u^\top A_\ell^\top A_\ell u = u^\top G_\ell u.$$

The first-order logit energy is $\|T_\ell u\|_2^2 = u^\top T_\ell^\top T_\ell u$, giving the displayed generalized Rayleigh quotient. Differentiating this quotient yields the generalized eigenproblem $T_\ell^\top T_\ell u = \lambda G_\ell u$.

For the subspace statement, set $w = G_\ell^{1/2} u$. Then $\|A_\ell u\|_2 = \|w\|_2$ and

$$T_\ell u = T_\ell G_\ell^{-1/2} w.$$

The k-dimensional average-energy problem under hidden-norm-isotropic perturbations is therefore the ordinary Ky Fan problem for $T_\ell G_\ell^{-1/2}$ in the w coordinates. Its solution is the top-k right singular subspace of $T_\ell G_\ell^{-1/2}$, mapped back by $G_\ell^{-1/2}$. □

L. Extended experimental results

This appendix provides complete tables, ablations, and supporting analyses summarized in Section 5.

L.1. Subspace-level forecasting in the controlled transformer

Setup. We train a 4-layer decoder-only transformer on the 26-token synthetic next-token task (Section 5) and extract activations and logits at forecast training steps $t_0 \in \{400, 1000, 2000\}$ and the final training step $t_1 = 20{,}000$. Per-layer pulled-back operators $\hat{T}_\ell = \hat{J}_\ell \hat{B}_\ell \hat{\Sigma}^{1/2}$ are estimated by ridge ($\hat{B}_\ell$ from features to residual stream, $\hat{J}_\ell$ from residual stream to logits); all predictions are fixed before later-checkpoint probes are evaluated.

Table 3. Steering benchmark in the controlled transformer, $n = 16$ seeds per family (4 directions $\times$ 4 perturbation magnitudes per seed, averaged within seed; 95% CI). Bold marks the family with the highest value in each row.

Metric	singular	mean-difference	probe	random
$\|\Delta\eta\|_2/\|\delta h\|_2$	$\mathbf{0.129 \pm 0.004}$	0.104 ± 0.004	0.037 ± 0.003	0.041 ± 0.002
Selectivity (target change / total logit-change norm)	$\mathbf{0.76 \pm 0.03}$	0.41 ± 0.09	0.55 ± 0.06	0.03 ± 0.06
Linearity ($\cos\angle(\Delta\eta,\ \hat{J}_\ell\,\delta h)$)	$\mathbf{0.78 \pm 0.03}$	0.50 ± 0.10	0.42 ± 0.07	0.06 ± 0.05

Calibration. Conjunction features over Bernoulli primitives have variances $\hat{\nu}_r = p_i(1-p_i)\,p_j(1-p_j)$ of order 10^{-2}. Under strict max calibration, where the logit clip is set to $\max_x |\eta_*(x)|$, the teacher distribution is nearly uniform ($H(p_*) \approx \log 32$), leaving only ≈ 0.03 nats of task-relevant signal. We therefore use p_{95} calibration in the main controlled-transformer experiment. The clip bound is set so that the 95-th percentile of $|\eta_*|$ equals the clip, giving $H(p_*) \approx 2.88$ nats and 0.59 nats of task-relevant signal. A small fraction of samples exceeds the bound and contributes to the κ_B prefactor in Theorem 3.1.

Early-to-late subspace preservation. The subspace-level prediction of the theory is that the top-k right singular subspace of the early $\hat{T}_\ell$ predicts the top-k subspace at the final checkpoint. We compute the principal-angle overlap $\Omega_k = \frac{1}{k}\,\|P_{V_k^{\text{early}}} V_k^{\text{late}}\|_F^2$ where V_k^{late} is the top-k right singular subspace of the ridge-fit pulled-back operator on the final-checkpoint cached activations and V_k^{early} is the same quantity at forecast training step $t_0 = 2000$. At the reference configuration (p_{95} calibration, layer 2, training step $t_0 = 2000$, $n = 16$ seeds), $\Omega_4 = \mathbf{0.74 \pm 0.07}$ versus a random-subspace baseline $k/m = 4/24 \approx 0.17$. The complete sweep across layers, forecast checkpoints, and k is in Appendix M.3; the future-only coordinate-masked variant and the hidden-space analogue are in Appendices M.4 and M.5.

L.2. Hidden-state-normalized steering benchmark

With the transformer frozen, we run the steering benchmark at the query position of the chosen layer, comparing singular, probe, mean-difference, and random direction families.

Theorem K.10 predicts that top singular directions of the pulled-back $\hat{T}_\ell$ maximize average first-order logit energy $\mathbb{E}\|\hat{T}_\ell u\|_2^2$ per unit-norm whitened feature-coordinate perturbation. We compare four direction families at layer $\ell = 2$, rank $k = 4$, perturbation magnitudes $\varepsilon \in \{0.25, 0.5, 1.0, 2.0\}$, across 16 seeds of the 4-layer transformer. For the selectivity metric, each steering direction is paired with its associated target direction in logit space. Singular directions use the corresponding logit-space left singular vector u_i of $\hat{T}_\ell$. Feature-indexed probe and mean-difference directions use the teacher output direction q_r for that feature. Each family contributes k directions.

Singular directions dominate on all three metrics with non-overlapping 95% CIs (Table 3). The first row is a hidden-state-norm analogue of Theorem K.10: the theorem's optimality statement is for perturbations in the whitened feature-coordinate norm, while Proposition K.11 gives the exact generalized-eigenvalue form for hidden Euclidean norm. In the experiment we map feature-space singular directions through $\hat{B}_\ell\hat{\Sigma}^{1/2}$ to hidden-state directions and normalize them in hidden-state Euclidean norm. Because the norm changes, this row should be interpreted as an empirical hidden-norm analogue rather than a direct test of Theorem K.10 under its original norm. The second row shows that the response is the most concentrated on the intended mode; the third shows that it tracks the local linear prediction $\hat{J}_\ell\,\delta h$ most faithfully.

L.3. Prospective validation in pretrained language models

The synthetic teacher–student and controlled transformer benchmarks test the theory in settings where feature values, teacher logits, and training dynamics are known by construction. Our pretrained-LM test asks whether the same object – the pulled-back operator $\hat{T}_\ell$ – carries predictive information in a pretrained Pythia model and on a natural-language agreement distinction.

We design a subject–verb agreement (SVA) benchmark with attractors. Each example is a noun phrase of the form "The *subj prep* the *attr*" with 20 morphologically regular (singular, plural) subject/attractor pairs, four prepositional templates, and seven target verb triples (*be, have, do, run, walk, know, love*); the task is to prefer the number-agreeing verb form at the next-token position (e.g. *is* vs *are*). Each example is tagged with three causal features (subject number, attractor number, their interaction) and eleven nuisance indicators (four template, seven verb lemma), with a lexical-plus-template held-out split (four subjects and one template withheld; 1,344 training and 896 test examples). The signal is the last-token margin

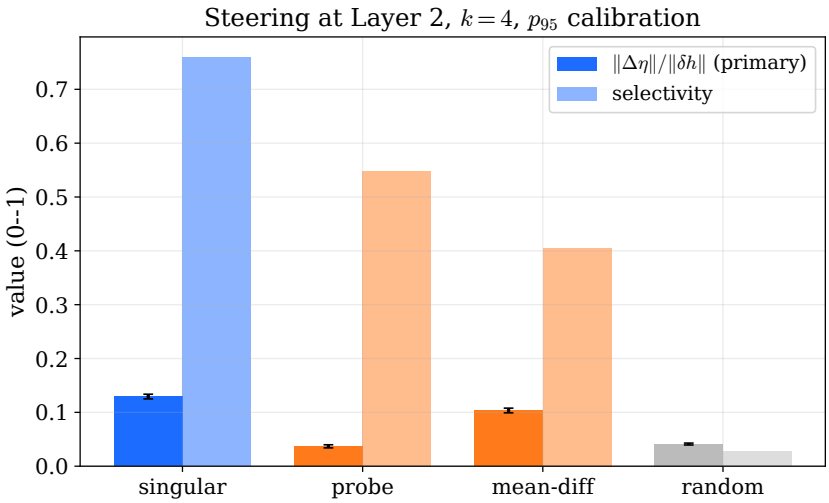

Figure 4. Reference-configuration steering results for layer 2, $k = 4$, and 95th-percentile calibration on the 4-layer transformer. Across 16 seeds, singular directions achieve higher logit-change gain and selectivity than probe, mean-difference, and random baselines.

Table 4. Final-checkpoint held-out margin-gradient-energy capture on SVA. Values are averaged over $n = 8$ seeds at each model's validation-selected layer; the last column reports test-margin accuracy.

Model	pulled-back D_4	random-4	hidden-PCA$_4$	grad-PCA$_4$ (oracle)	test-margin acc
Pythia-70M, L_6	0.278	0.008	0.172	0.844	0.94
Pythia-160M, L_{12}	0.300	0.005	0.262	0.851	0.92
Pythia-410M, L_{23}	0.221	0.004	0.161	0.730	0.92

$m(x) = \log p(\text{correct}) - \log p(\text{incorrect})$ and its gradient $g_\ell(x) = \nabla_{h_\ell} m(x)$ at layer ℓ at the same token.

Construction of $\hat{T}_\ell$. We fit $\hat{B}_\ell : \phi \mapsto h_\ell$ on the train split by ridge regression, form the whitened feature covariance $\hat{\Sigma}^{1/2}$ on train, and stack per-example margin gradients into $G_{\text{tr}} \in \mathbb{R}^{N_{\text{tr}} \times d}$. The pulled-back operator is

$$\hat{T}_\ell = G_{\text{tr}}\, \hat{B}_\ell\, \hat{\Sigma}^{1/2} \;\in\; \mathbb{R}^{N_{\text{tr}} \times m}.$$

Its top-k right singular vectors $V_k \in \mathbb{R}^{m \times k}$ map via $D_k = \text{orth}(\hat{B}_\ell \hat{\Sigma}^{1/2} V_k)$ to residual-stream directions predicted to carry the task-relevant interaction structure. All quantities are estimated on train only; all scoring is on test.

L.3.1. Held-out forecast accuracy on margin gradients

We ask what fraction of the test-set margin-gradient energy $\|G_{\text{te}}\|_F^2$ is captured by the top-k directions D_k, compared with three baselines: (i) a random-k subspace; (ii) top-k PCA of the train residual stream H_{tr}; and (iii) top-k PCA of the train gradients G_{tr}, a gradient-PCA reference baseline trained on the same gradient quantity. We run Pythia-70M, Pythia-160M, and Pythia-410M at six revisions (step $\{1, 4, 16, 48, 96, 143\} \cdot 10^3$) with 4 random seeds per (model, revision) for non-final revisions, and a second 4-seed batch at the final revision ($n = 8$ seeds per (model, final-revision) configuration).

At the final checkpoint and each model's validation-selected layer (L_6, L_{12}, L_{23}, respectively, chosen by held-out pulled-back gradient-energy capture at k=4) with $k = 4$ directions out of hidden dimensions $\{512, 768, 1024\}$, the pulled-back direction captures a substantial fraction of test-gradient energy: All values are averaged across the $n = 8$ seeds at the final revision. Pulled-back is 35–60$\times$ the random baseline at every model size and every final-checkpoint configuration. At the final checkpoint the gap to hidden-PCA shrinks as the task saturates (Pythia-160M: 0.300 vs 0.262); the advantage that distinguishes pulled-back from hidden-PCA is at the *early* checkpoints, where the forecast is most informative. Section L.3.2 reports the early-to-late results.

The forecast gap is largest at early checkpoints. At step 4000, when each model has reached only 84–93% test-margin accuracy, pulled-back already identifies test-gradient directions that hidden-PCA cannot:

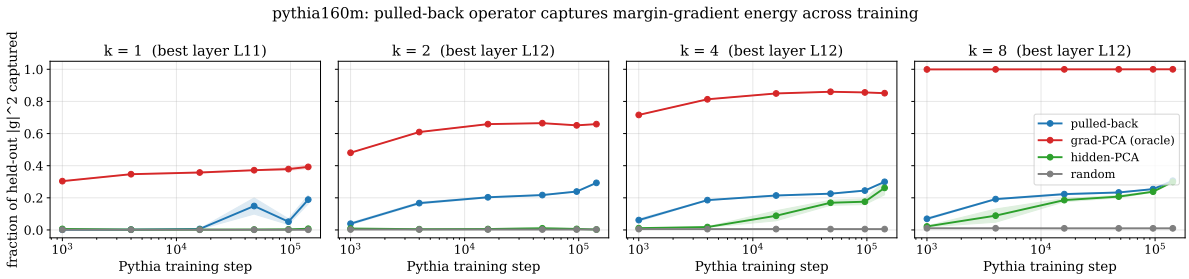

Figure 5. Pythia-160M subject–verb agreement at the model's best layer L_{12}: fraction of held-out margin-gradient energy captured by the top-k pulled-back directions (blue), compared with random (gray), hidden-PCA (green), and gradient-PCA oracle (red), across Pythia revisions. Shaded bands denote 95% CIs across 4 seeds. Figures for the other two model sizes are in Appendix M.8.

Table 5. Early-to-late SVA forecast. The pulled-back subspace is fit at early checkpoint t_0, frozen, and evaluated against final-checkpoint held-out margin-gradient energy at t_1 = step 143,000.

configuration	$D^{t_0} \to t_1$ (ours)	final pulled-back D^{t_1}	$\mathrm{PCA}_h^{t_0}$	$\mathrm{PCA}_g^{t_0}$	overlap Ω_k
Pythia-70M, t_0 = step 4,000	0.150 ± 0.003	0.276	0.022 ± 0.003	0.383	0.19
Pythia-70M, t_0 = step 16,000	0.190 ± 0.002	0.276	0.052 ± 0.022	0.483	0.27
Pythia-70M, t_0 = step 48,000	0.224 ± 0.006	0.276	0.072 ± 0.014	0.613	0.37
Pythia-160M, t_0 = step 4,000	0.176 ± 0.001	0.302	0.012 ± 0.002	0.411	0.27
Pythia-160M, t_0 = step 16,000	0.173 ± 0.005	0.302	0.065 ± 0.024	0.549	0.34
Pythia-160M, t_0 = step 48,000	0.197 ± 0.003	0.302	0.141 ± 0.014	0.677	0.39
Pythia-410M, t_0 = step 4,000	0.241 ± 0.006	0.222	0.160 ± 0.019	0.415	0.41
Pythia-410M, t_0 = step 16,000	0.214 ± 0.006	0.222	0.015 ± 0.001	0.567	0.66
Pythia-410M, t_0 = step 48,000	0.229 ± 0.007	0.222	0.053 ± 0.029	0.664	0.80

Model (step 4000)	pulled-back D_4	hidden-PCA_4	ratio
Pythia-70M, L_6	0.164	0.024	6.8×
Pythia-160M, L_{12}	0.186	0.018	10.3×
Pythia-410M, L_{23}	0.176	0.119	1.5×

This is the regime in which the theory is most informative: at early checkpoints the task is not yet saturated, so hidden-PCA, which is dominated by pretraining-era variance, carries little task-specific signal, while the pulled-back operator already captures the directions that will eventually encode the task. Complete emergence curves are in Figure 5 and Appendix M.8.

L.3.2. Early-to-late forecasting

To separate the same-checkpoint explanation of the margin-gradient structure from a true early-checkpoint forecast, we freeze the top-k residual-stream subspace $D_\ell^{t_0}$ computed from the pulled-back operator at an early checkpoint t_0, then score that frozen subspace against margin-gradient test energy at the final checkpoint t_1 = step 143,000. Baselines at the same early checkpoint are hidden-PCA on the train residual stream and gradient-PCA on the train margin gradients. At every model/checkpoint configuration (model, t_0) with $t_0 \in$ {step 4,000, 16,000, 48,000} and $k = 4$, the frozen early pulled-back subspace captures more late-checkpoint test-gradient energy than the frozen early hidden-PCA subspace: The frozen early pulled-back subspace captures 1.4–14.6× as much held-out test-gradient energy as early hidden-PCA. It is within 0.04 to 0.13 of the final pulled-back reference in every configuration. At Pythia-410M with t_0 = step 4,000, the frozen early capture fraction is within the 4-seed sampling variation of the late-oracle value. The principal-angle overlap between early and late top-4 right singular subspaces grows from 0.19 (step 4,000, 70M) to 0.80 (step 48,000, 410M), consistent with the subspace-level preservation theorem.

L.3.3. Replication on reflexive agreement

To check that the SVA result is not specific to the subject–verb number-agreement task, we repeat the analysis on an English reflexive agreement benchmark with an attractor. Each example has the form "The *subj* that the *attr* hired told {himself, themselves}", with an analogous feature dictionary (subject number, attractor number, their interaction, four

Table 6. Reflexive-agreement replication at the SVA-selected layer and $k = 4$. Pulled-back directions are compared with hidden-PCA, random subspaces, and a gradient-PCA oracle on the held-out test partition.

configuration (reflexive)	pulled-back	hidden-PCA	random	grad-PCA	test acc
Pythia-70M, L_6	0.371 ± 0.017	0.008 ± 0.004	0.007 ± 0.001	1.000	0.81
Pythia-160M, L_{12}	0.276 ± 0.027	0.058 ± 0.016	0.005 ± 0.001	1.000	0.95

templates; $m = 7$). We evaluate at the SVA-selected layer, with $k = 4$, on the test partition of the lexical-plus-template held-out split: The pulled-back-to-hidden-PCA ratios are $46\times$ on Pythia-70M and $4.8\times$ on Pythia-160M at the same-checkpoint final revision, confirming that the gradient-energy forecast advantage is not specific to the subject–verb number-agreement task.

Reflexive-agreement early-to-late forecast. On the same benchmark we also run the early-to-late protocol (frozen early pulled-back subspace evaluated on final-checkpoint margin-gradient energy):

config (reflexive)	pulled-back	hidden-PCA	grad-PCA
Pythia-70M, $t_0 = 4{,}000$	0.172 ± 0.009	0.097 ± 0.049	0.405
Pythia-70M, $t_0 = 16{,}000$	0.210 ± 0.003	0.005 ± 0.002	0.582
Pythia-70M, $t_0 = 48{,}000$	0.210 ± 0.013	0.009 ± 0.006	0.806
Pythia-160M, $t_0 = 4{,}000$	0.219 ± 0.012	0.143 ± 0.052	0.670
Pythia-160M, $t_0 = 16{,}000$	0.225 ± 0.012	0.134 ± 0.038	0.778
Pythia-160M, $t_0 = 48{,}000$	0.270 ± 0.008	0.121 ± 0.069	0.889

Ratios of pulled-back to hidden-PCA performance are $1.5\times$–$42\times$ across configurations. The method therefore transfers to a second interaction task in the same structural family without task-specific tuning.

L.4. Robustness checks

Complete (ℓ, k) steering sweep. Beyond the reference configuration of Table 3, we run the evaluation protocol across $\ell \in \{1, 2, 3, 4\}$ and $k \in \{1, 2, 4, 8, 12\}$ on the 4-layer transformer (16 seeds each, p_{95} calibration). Appendix M.1 reports the complete grid of per-configuration means and CIs. The singular family is best on the three reported metrics in 19 of 20 (ℓ, k) configurations. Its mean fidelity ratio is 6.57 relative to probe (min 1.04, max 66.27) and 1.34 relative to mean-difference (min 0.79, max 2.04).

Matched-perturbation Pareto. To isolate per-unit-perturbation performance from differences in total logit change, we interpolate each family's ε-sweep onto a common Δ_{total} axis and compare target lift at matched Δ_{total}. Across the 20 (ℓ, k) configurations, the singular family gives $\mathbf{1.25}\times$ the target lift of the probe family and $\mathbf{1.75}\times$ the target lift of the mean-difference family at matched total change (ranges 0.83–1.94 and 1.01–4.41 respectively). The per-unit-perturbation advantage is therefore not an artifact of larger $\|\delta h\|$.

Tail-linearity diagnostic and steering fidelity. We test whether lower empirical tail distortion is associated with higher local-linear steering fidelity. Across the 48 matched (seed, ℓ) configurations at $k = 4$ from 16 seeds, the Spearman correlation between $\hat{\delta}_\ell / \|\hat{J}_\ell h\|_{\text{RMS}}$ and the singular-family fidelity cosine is $\rho = \mathbf{-0.92}$ with a seed-clustered percentile bootstrap 95% CI of $[-0.94, -0.89]$ over 2000 resamples; every bootstrap replicate returned a negative correlation. The mean-difference family shows a similar association ($\rho = -0.89$, 95% CI $[-0.94, -0.81]$), while the probe family gives a weakly positive association ($\rho = +0.33$, CI $[+0.10, +0.55]$) which does not match the prediction.

M. Additional experimental tables and figures

This appendix collects supporting experimental tables and figures that complement the main results in Section 5. The subsections below give the complete steering sweep, causal-intervention tables, controlled-transformer subspace-overlap variants, cross-model feature-space forecasts, and the pretrained-LM emergence curves. These results are reported separately to keep the main experimental narrative focused while preserving the complete quantitative record.

M.1. Complete (ℓ, k) steering sweep

We sweep $\ell \in \{1, 2, 3, 4\}$ and $k \in \{1, 2, 4, 8, 12\}$ on the 4-layer transformer under p_{95} calibration, $n = 16$ seeds per (ℓ, k) configuration. Table 7 reports, for each (ℓ, k), the ratio of the singular-family fidelity cosine to each baseline fidelity cosine. Values above 1 indicate higher fidelity for the singular family. The singular family's fidelity is $\geq$ each baseline at 19 of 20 configurations (the ratio relative to mean-difference is 0.79 at $\ell = 1, k = 1$). Averaged across configurations, the ratios are 1.34 for singular versus mean-difference (min 0.79, max 2.04), 6.57 for singular versus probe (min 1.04, max 66.27), and 11.18 for singular versus random (min 2.93, max 47.42). The singular family's $\|\Delta\eta\|_2/\|\delta h\|_2$ increases with depth.

Table 7. Complete (ℓ, k) steering sweep over 16 seeds per configuration. The table reports the singular-family logit-change gain and its fidelity-cosine ratio relative to mean-difference, probe, and random baselines. Ratios exceed 1 at 19 of 20 configurations; the exception is the mean-difference comparison at $\ell = 1, k = 1$.

ℓ	k	$\|\Delta\eta\|/\|\delta h\|$	sing/md	sing/pr	sing/rand
1	1	0.122	0.79	1.04	2.93
1	2	0.124	1.60	2.05	8.02
1	4	0.120	2.03	2.19	11.26
1	8	0.113	2.04	1.70	7.60
1	12	0.110	1.46	1.37	4.68
2	1	0.140	1.22	1.61	4.85
2	2	0.137	1.32	1.78	9.03
2	4	0.129	1.55	1.88	13.71
2	8	0.123	1.95	1.78	9.06
2	12	0.117	1.82	1.69	9.47
3	1	0.160	1.02	2.15	7.98
3	2	0.149	1.08	2.33	12.76
3	4	0.137	1.20	2.33	19.88
3	8	0.130	1.32	2.60	47.42
3	12	0.124	1.28	2.86	18.84
4	1	0.162	1.00	1.80	8.36
4	2	0.154	1.00	3.06	7.29
4	4	0.140	1.00	11.69	6.57
4	8	0.129	1.08	19.27	7.14
4	12	0.128	1.05	66.27	6.74

M.2. Complete causal-intervention tables

This subsection reports the per-pair DiD statistics summarized in Section 5.3. At every bottleneck width k, the higher-ranked-minus-lower-ranked R^2 advantage is computed separately in condition A and condition B, and then summed across conditions to form the paired difference-in-differences. A commonness-only or usefulness-only ranking must disagree with the product rule in exactly four of the eight pairs, so a non-product ranking is expected to give near-zero summed advantage for the pair family it disagrees with. The product rule selects the higher-R^2 feature in both conditions in every pair at every width at or above the emergence frontier $k = 4$.

M.3. Subspace-overlap tables

We report the principal-angle overlap $\Omega_k = \frac{1}{k}\|P_{V_k^{\text{early}}} V_k^{\text{late}}\|_F^2$ between the top-k right singular subspaces at forecast training step $t_0 = 2000$ and final training step t_1. At the reference configuration (layer 2, training step $t_0 = 2000$, $k = 4$, p_{95} calibration, $n = 16$ seeds): $\Omega_4 = 0.74 \pm 0.07$ versus the random baseline $k/m = 4/24 \approx 0.17$. The complete sweep over layers, t_0, and k is:

layer ($t_0 = 2000$)	$k=1$	$k=2$	$k=4$	$k=8$
layer 0 (embed)	1.000±0.000	0.631±0.056	0.798±0.034	0.870±0.021
layer 1	0.821±0.099	0.832±0.073	0.829±0.065	0.798±0.052
layer 2	0.775±0.175	0.823±0.082	0.743±0.072	0.836±0.028
layer 3	0.987±0.007	0.935±0.046	0.732±0.047	0.868±0.028
layer 4	0.991±0.001	0.982±0.005	0.753±0.044	0.914±0.017
chance k/m	0.042	0.083	0.167	0.333

Table 8. Per-pair causal-intervention results. The table reports higher-ranked-minus-lower-ranked R^2 advantages in conditions A and B, their difference-in-differences (DiD) sum, and seed-bootstrap 95% CIs across 8 seeds. The maximum possible DiD sum is $+2.0$.

pair	type	k	advantage A	advantage B	DiD sum
0	usefulness-flip	2	$+0.003 \pm 0.002$	$+0.001 \pm 0.001$	$+0.004 \pm 0.002$
0	usefulness-flip	4	$+0.010 \pm 0.006$	$+0.019 \pm 0.013$	$+0.029 \pm 0.016$
0	usefulness-flip	8	$+0.967 \pm 0.005$	$+0.950 \pm 0.013$	$+1.917 \pm 0.015$
1	usefulness-flip	2	$+0.016 \pm 0.015$	$+0.009 \pm 0.007$	$+0.025 \pm 0.016$
1	usefulness-flip	4	$+0.915 \pm 0.059$	$+0.888 \pm 0.046$	$+1.803 \pm 0.093$
1	usefulness-flip	8	$+0.972 \pm 0.010$	$+0.974 \pm 0.008$	$+1.946 \pm 0.014$
2	usefulness-flip	2	$+0.949 \pm 0.016$	$+0.964 \pm 0.010$	$+1.913 \pm 0.016$
2	usefulness-flip	4	$+0.967 \pm 0.013$	$+0.971 \pm 0.007$	$+1.938 \pm 0.015$
2	usefulness-flip	8	$+0.978 \pm 0.005$	$+0.975 \pm 0.004$	$+1.953 \pm 0.008$
3	usefulness-flip	2	$+0.976 \pm 0.007$	$+0.978 \pm 0.005$	$+1.954 \pm 0.010$
3	usefulness-flip	4	$+0.978 \pm 0.006$	$+0.981 \pm 0.004$	$+1.959 \pm 0.009$
3	usefulness-flip	8	$+0.976 \pm 0.005$	$+0.980 \pm 0.005$	$+1.955 \pm 0.004$
4	commonness-flip	2	$+0.007 \pm 0.006$	$+0.004 \pm 0.006$	$+0.011 \pm 0.009$
4	commonness-flip	4	$+0.813 \pm 0.070$	$+0.777 \pm 0.117$	$+1.590 \pm 0.163$
4	commonness-flip	8	$+0.968 \pm 0.005$	$+0.963 \pm 0.013$	$+1.930 \pm 0.011$
5	commonness-flip	2	$+0.011 \pm 0.007$	$+0.003 \pm 0.003$	$+0.014 \pm 0.006$
5	commonness-flip	4	$+0.085 \pm 0.046$	$+0.079 \pm 0.041$	$+0.164 \pm 0.066$
5	commonness-flip	8	$+0.970 \pm 0.011$	$+0.969 \pm 0.006$	$+1.939 \pm 0.014$
6	commonness-flip	2	$+0.008 \pm 0.003$	$+0.009 \pm 0.004$	$+0.017 \pm 0.007$
6	commonness-flip	4	$+0.123 \pm 0.071$	$+0.179 \pm 0.163$	$+0.302 \pm 0.203$
6	commonness-flip	8	$+0.973 \pm 0.007$	$+0.968 \pm 0.010$	$+1.941 \pm 0.013$
7	commonness-flip	2	$+0.003 \pm 0.005$	$+0.002 \pm 0.004$	$+0.005 \pm 0.004$
7	commonness-flip	4	$+0.028 \pm 0.008$	$+0.014 \pm 0.013$	$+0.042 \pm 0.013$
7	commonness-flip	8	$+0.953 \pm 0.013$	$+0.932 \pm 0.020$	$+1.885 \pm 0.029$

Preservation is well above chance at every non-embedding layer and every $k \in \{1, 2, 4, 8\}$.

M.4. Future-masked subspace overlap

To test the forecast on the *not-yet-emerged* features, we mask the already-emerged feature coordinates (probe $R^2 \geq 0.8$ at t_0) and measure principal-angle overlap of $V_k(t_0)$ vs $V_k(t_1)$ after restriction to the complement. Let r_0 be the number of emerged features at t_0; the chance baseline is $k/(m - r_0)$. The following values use $n = 16$ seeds and p_{95} calibration.

Table 9. Future-masked subspace overlap at the controlled-transformer reference layer. The table reports overlap after removing already readable coordinates; r_0 is the mean number removed, and CIs are seed-clustered percentile bootstrap intervals over 16 seeds. The chance baseline is $k/(m - r_0)$.

layer, t_0	r_0	Ω_4^{fut}	random k	chance
layer 2, $t_0 = 400$	2.2	0.351 ± 0.046	0.191 ± 0.007	0.184
layer 2, $t_0 = 1000$	12.1	0.507 ± 0.069	0.345 ± 0.035	0.347
layer 2, $t_0 = 2000$	15.1	0.66 ± 0.04	0.493 ± 0.078	0.497
layer 3, $t_0 = 2000$	21.6	0.812 ± 0.076	0.755 ± 0.063	0.756

At the reference configuration (layer 2, training step $t_0 = 2000$, $k = 4$), the future-only overlap is $\mathbf{0.66 \pm 0.04}$ versus average chance 0.497 and a random-subspace baseline of 0.493 ± 0.078. Because r_0 varies across seeds, the chance baseline is itself random; comparing per-seed gives $\Omega_4^{\mathrm{future}} -$ chance $= 0.170 \pm 0.057$ over 16 seeds (with $13/16$ seeds having $\Omega_4^{\mathrm{future}} >$ chance). Deeper layers where most features have already emerged (layer 3 at training step $t_0 = 2000$ has $r_0 = 21.6$) are close to chance because the residual dimension $m - r_0$ is small.

Robustness to the emergence threshold. The reference configuration above uses the paper-wide emergence rule (probe $R^2 \geq 0.8$). At tighter thresholds the predicted not-yet-emerged subspace is larger (smaller r_0), and $\Omega_4^{\mathrm{future}} -$ chance grows. At the same reference cell, $n = 16$ seeds:

Table 10. Cross-model subspace transfer on SVA. Early pulled-back subspaces from smaller Pythia models are compared with the final-checkpoint pulled-back subspace of the larger model in the shared whitened feature space.

small → big	t_0	small→big Ω_4	small-final→big-final	random	chance
70M → 160M	step 4,000	0.660 ± 0.013	0.720	0.273	0.286
70M → 160M	step 16,000	0.642 ± 0.031	0.720	0.273	0.286
70M → 160M	step 48,000	0.676 ± 0.022	0.720	0.273	0.286
70M → 410M	step 4,000	0.727 ± 0.013	0.758	0.276	0.286
160M → 410M	step 4,000	0.749 ± 0.022	0.762	0.276	0.286

threshold	r_0	Ω_4^{fut}	chance	per-seed gap	positive seeds
$R^2 \geq 0.80$	15.1	0.666 ± 0.040	0.497	0.170 ± 0.057	13/16
$R^2 \geq 0.85$	13.4	0.619 ± 0.046	0.409	0.210 ± 0.064	14/16
$R^2 \geq 0.90$	12.1	0.599 ± 0.041	0.351	$\mathbf{0.248 \pm 0.047}$	**16/16**
$R^2 \geq 0.95$	8.9	0.575 ± 0.043	0.272	$\mathbf{0.304 \pm 0.045}$	**16/16**

At thresholds ≥ 0.90, every seed has a positive gap. The gap-to-CI ratio rises from 3.0 at the paper-wide 0.8 threshold to 6.8 at 0.95. Thus, the predicted subspace exceeds chance even when restricted to not-yet-emerged coordinates, and the evidence strengthens under stricter emergence criteria.

M.5. Hidden-space early-to-late overlap in the controlled transformer

We run the controlled-transformer analogue of the Pythia early-to-late experiment (Appendix L.3.2). For each (seed, layer, t_0, k), we compute the hidden-space top-k direction $D_k(t) = \mathrm{orth}(\hat{B}_\ell(t)\hat{\Sigma}^{1/2}(t)V_k(t))$ at both the early forecast checkpoint and the final step; the reported quantity is $\Omega_k^{\mathrm{hidden}}(D_k(t_0), D_k(t_1))$. The baselines are hidden-PCA of $H_\ell(t_0)$ and a random k-subspace. Reference-cell values (layer 2, training step $t_0 = 2000$, $k = 4$, $n = 16$ seeds, p_{95} calibration):

family	overlap $\Omega_4^{\mathrm{hidden}}$
pulled-back $D_k(t_0)$	$\mathbf{0.671 \pm 0.064}$
hidden-PCA at t_0	0.372 ± 0.055
random k-subspace	0.021 ± 0.001

The pulled-back early direction overlaps the late pulled-back direction at $1.8\times$ the hidden-PCA overlap and $32\times$ the random overlap. The same pattern holds at layers 3 and 4 at training step $t_0 = 2000$, with $\Omega_4^{\mathrm{hidden}} = 0.671 \pm 0.047$ and 0.716 ± 0.043 respectively, and is monotone in k at deeper layers.

M.6. Cross-model feature-space forecast

We compare top-k right singular subspaces of $\hat{T}_\ell$ in whitened feature space across Pythia sizes. Each model uses its validation-selected layer (L_6, L_{12}, L_{23} for 70M/160M/410M) and the 14-dimensional SVA feature dictionary is shared across sizes. For each small-model early checkpoint $t_0 \in \{\text{step } 4{,}000,\ \text{step } 16{,}000,\ \text{step } 48{,}000\}$ we measure the principal-angle overlap to the larger model's final-checkpoint subspace, averaging over seeds. The early feature-space subspace in a smaller model overlaps the larger model's final-checkpoint feature subspace at $\Omega_4 = 0.66$–0.75 vs chance 0.29 and random 0.27, and is close to the upper bound given by the seed-averaged small-final to big-final overlap.

M.7. Predicted vs empirically readable subspace

For each run we compute the empirical readability operator on held-out data,

$$\hat{R}_\ell = \hat{\Sigma}_{\tilde{\phi}h}\left(\hat{\Sigma}_{hh} + \lambda_h I\right)^{-1}\hat{\Sigma}_{h\tilde{\phi}},$$
$$\tilde{\phi} = \hat{\Sigma}^{-1/2}\phi, \quad \lambda_h = 10^{-3}\,\frac{\mathrm{tr}(\hat{\Sigma}_{hh})}{d},$$

We then report the principal-angle overlap between its top-k eigenspace and the top-k right singular subspace of $\hat{T}_\ell$ in whitened feature space. This tests whether the k-dimensional feature subspace selected by the theory is also the subspace most readable by ridge probes from the residual stream. Results at the final checkpoint of the 4-layer controlled transformer (p_{95}-calibrated, $n = 16$ seeds, per layer and per k, $\pm 95\%$ CI):

layer	$k=1$	$k=2$	$k=4$	$k=8$
embed (L_0)	0.054±0.036	0.119±0.036	0.189±0.022	0.339±0.019
after L_1	0.171±0.130	0.312±0.106	0.615±0.062	0.836±0.035
after L_2	0.192±0.167	0.324±0.119	0.511±0.069	0.581±0.050
after L_3	0.146±0.155	0.276±0.112	0.460±0.068	0.667±0.052
after L_4	0.249±0.161	0.301±0.085	0.426±0.057	0.756±0.029
random k/m	0.042	0.083	0.167	0.333

The embedding row is at the random baseline, as expected. All non-embedding configurations at $k \in \{4, 8\}$ are 2.3–3.7× the corresponding random baseline. At the reference configuration (L_2, $k = 4$) the overlap is 0.51 ± 0.07.

M.8. Pretrained-LM emergence curves

Figures 6 and 7 give the Pythia-70M and Pythia-410M emergence curves at $k \in \{1, 2, 4, 8\}$ and their respective best layers. At every size, the pulled-back curve remains above hidden-PCA at every checkpoint at or before step 16,000, with non-overlapping 95% CIs across those checkpoints; the hidden-PCA curve rises later and approaches pulled-back only by step 143,000.

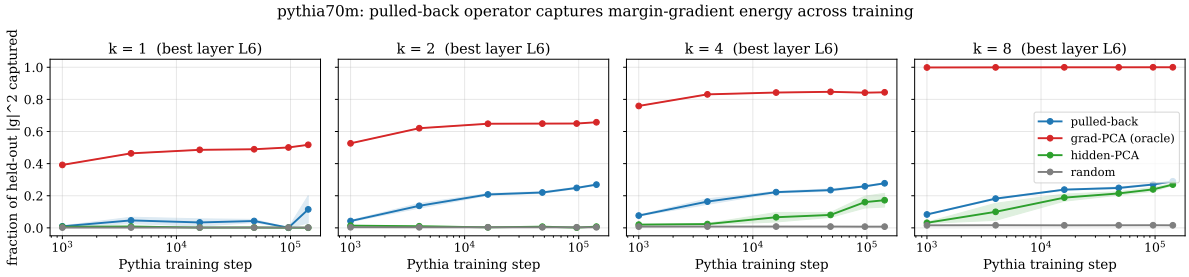

Figure 6. Pythia-70M at L_6 on the SVA benchmark: fraction of held-out margin-gradient energy captured by the top-k pulled-back directions (blue), compared with random, hidden-PCA, and gradient-PCA oracle baselines across six Pythia revisions. Shaded bands denote 95% CIs across 4 seeds.

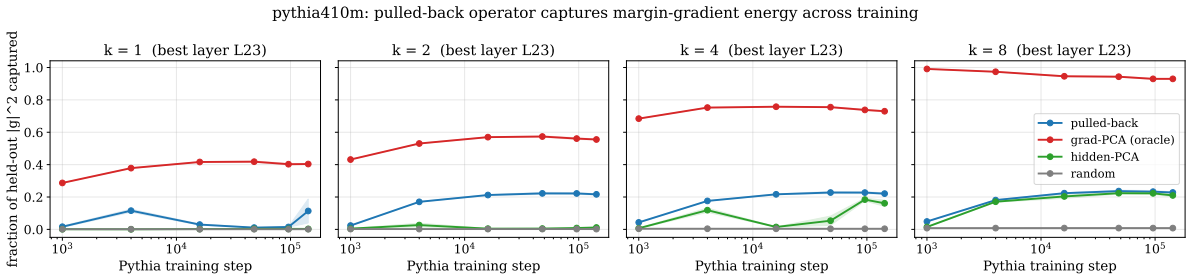

Figure 7. Pythia-410M at L_{23} on the SVA benchmark: fraction of held-out margin-gradient energy captured by the top-k pulled-back directions, compared with random, hidden-PCA, and gradient-PCA oracle baselines across six Pythia revisions. Shaded bands denote 95% CIs across 4 seeds.